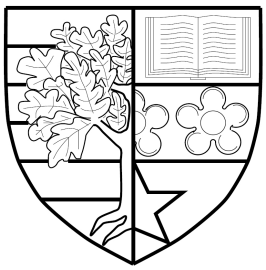
AUTOMATIC HUMAN BEHAVIOUR ANOMALY DETECTION IN SURVEILLANCE VIDEO

by
Michael Jeremy Vincent Leach



Submitted for the degree of
Doctor of Engineering

Institute of Signal Sensor Systems
School of Engineering and Physical Sciences
Heriot-Watt University

February 2015

Abstract

This thesis work focusses upon developing the capability to automatically evaluate and detect anomalies in human behaviour from surveillance video. We work with static monocular cameras in crowded urban surveillance scenarios, particularly airports and commercial shopping areas. Typically a person is 100 to 200 pixels high in a scene ranging from 10 - 20 meters width and depth, populated by 5 to 40 people at any given time. Our procedure evaluates human behaviour unobtrusively to determine outlying behavioural events, flagging abnormal events to the operator.

In order to achieve automatic human behaviour anomaly detection we address the challenge of interpreting behaviour within the context of the social and physical environment. We develop and evaluate a process for measuring social connectivity between individuals in a scene using motion and visual attention features. To do this we use mutual information and Euclidean distance to build a social similarity matrix which encodes the social connection strength between any two individuals. We develop a second contextual basis which acts by segmenting a surveillance environment into behaviourally homogeneous subregions which represent high traffic slow regions and queuing areas. We model the heterogeneous scene in homogeneous subgroups using both contextual elements. We bring the social contextual information, the scene context, the motion, and visual attention features together to demonstrate a novel human behaviour anomaly detection process which finds outlier behaviour from a short sequence of video. The method, Nearest Neighbour Ranked Outlier Clusters (NN-RCO), is based upon modelling behaviour as a time independent sequence of behaviour events, can be trained in advance or set upon a single sequence. We find that in a crowded scene the application of Mutual Information-based social context permits the ability to prevent self-justifying groups and propagate anomalies in a social network, granting a greater anomaly detection capability. Scene context uniformly improves the detection of anomalies in all the datasets we test upon.

We additionally demonstrate that our work is applicable to other data domains. We demonstrate upon the Automatic Identification Signal data in the maritime domain. Our work is capable of identifying abnormal shipping behaviour using joint motion dependency as analogous for social connectivity, and similarly segmenting the shipping environment into homogeneous regions.

Contents

List of Figures

List of Acronyms

AIS Automatic Identification System

EM Expectation-Maximization

TPR True Positive Rate

FPR False Positive Rate

HMM Hidden Markov Model

MLE Maximum Likelihood Estimation

MDL Minimum Description length

HOG Histogram of Oriented Gradients

R-HOG Rectangular Histogram of Oriented Gradients

GPS Global Positioning System

TLD Tracking Learning Detecting

DPM Deformable Part-Based Model

SVM Support Vector Machine

RSVM Robust Support Vector Machines

CNN Convolutional Neural Networks

WAMI Wide Area Motion Imagery

LK Lucas Kanade

PCA Principal Component Analysis

CMD Compact Matrix Decomposition

PETS Performance Evaluation of Tracking and Surveillance

MAE mean angular error

EMD Earth Mover Distance

CasDBN cascade of Dynamic Bayesian Network

MOHMM Multi-Observation Hidden Markov Model

DBN Dynamic Bayesian Network

GNSS Global navigation satellite system

VHF Very High Frequency

PDF Probability Density Function

GMM Gaussian Mixture Model

NN-RCO Nearest Neighbour - Ranked Cluster Outliers

QT Quad Tree

ATTAIN Accurate Target Tracking and Identification

TP True Positive

FP False Positive

Chapter 1

Introduction

As a society we have the need to monitor public and private space in order to prevent criminal behaviour and identify security threats. The scale at which surveillance is undertaken and the density of information in video results in a huge amount of data - the analysis of which using human resources is often prohibitively expensive. The solution is to automate human surveillance [63]. Automatic human behaviour anomaly detection is an endeavour to enable a computer to model human behaviour and detect outlier abnormal behaviour. This task entails many challenges, not least overcoming subtlety, obscurity, and the dependency upon prior knowledge. Human behaviour changes its meaning when seen in different contexts, and even human observers may disagree upon interpretation when observing the same data, adding the problem of subjectivity. Regardless, the prevalence of video surveillance and overwhelming quantity of data leads to the desire to automatically highlight salient and abnormal behaviour. It is this task that we hope to progress a solution towards in this thesis. We build upon the advances in machine learning and behaviour modelling which are complimented by recent advances in pedestrian detection and robust long term human tracking, all bringing us closer to autonomously profiling the individual behaviours within a crowded surveillance scene.

Early methods of supporting surveillance analysis and detecting abnormal human behaviour were based upon signature recognition that rely upon a-priori defined models of an abnormal behaviour. These, typically inflexible, template patterns may be defined by expert domain knowledge such as the methods of Edlund et al. [28], or trained upon observations of previous abnormal behaviour Fooladvandi et al. [31]. However, signature based methods do not scale in complexity as the complexity of the domain increases. This is due to the practical limitations of modelling all possible variations of behaviour, limited knowledge of potential behaviour classes, and limitations in encoding expert knowledge into a representative encompassing model. For that reason there is a requirement to develop data driven methods that learn the characteristics of normal and abnormal behaviour from exemplar data. The analysis should focus upon detecting statistically 'strange and abnormal' instances of data instead of pre-defined classes. The model of normality can be extracted from the training data, and abnormal data is identified by the appearance of being produced by a separate distribution than the normal training data. Such an anomaly detection method follows the paradigm of outlier detection. The complexity of the problem is greatly reduced from that of templates, as we no longer require a wealth of social, political, and cultural information to interpret the behaviour observed. But instead we rely upon the assumption that what is abnormal is rare.

The crux of anomaly detection methods is determining how to represent the

observable features in which the behaviours reside, and how to measure similarity between these representations. Events in the video, be they simple pixel motion or complex agent behaviours, must be encoded in such a way that the defining characteristics that separate normal behaviour from abnormal behaviour are harnessed. To determine outliers, a suitable metric must be devised to measure the distance between behaviours. The choice of representation and similarity metric determine the nature of the anomalies that can be detected. In a highly constrained, or behaviourally homogeneous, scene it may be enough to use the similarity of trajectories to detect novel behaviours through the scene. However, for more complex, dynamic, or behaviourally heterogeneous environments, further steps must be taken to ensure that there is distinction between separate classes. We follow from this paradigm in our approach. However, rather than merely defining a system that better fits the nuances of a particular environment or set of behaviours, we approach the problem by breaking the ambiguity that arises from behaviourally heterogeneous surveillance scenes by using contextual information. We tackle commonplace environments such as airports or urban shopping regions which are dynamic, changing over time and entailing different contexts in which the interpretation of behaviour is changed. We address this problem whilst finding abnormal behaviours as those that are statistically salient in comparison to other observed behaviours.

1.1 Motivation

Our motivation is threefold. Firstly, providing automation to surveillance would bring the analysis of human behaviour in surveillance closer to real time. Real time, automatic, human behaviour anomaly detection is a highly desired goal in this field as it will provide the ability to use surveillance data as preventative of crime, rather than merely a forensic source after a crime. To reach this goal we must first overcome hurdles in detection and tracking, speed and accuracy, and the depth and discriminative power of automated behaviour interpretation. Our second motivation is to enhance the existing capability of surveillance operators. Behavioural anomalies are naturally sparse and often behaviour is played out over long enough a time period that the sense of continuity is lost to the observer, particularly when observing multiple targets. The strengths of an automated interpretation of behaviour is that such as system can handle long observations without dropping attention, has no prior prejudices, and can monitor multiple targets with equal scrutiny. Our final motivation for this work is to provide algorithms which are applicable in alternative domains, particular the maritime domain, largely due to our close working relationship with the defence industry. Maritime behaviour analysis provides an environment in which the tracking is relatively simple and the accuracy is high. In the maritime domain the behaviour is more constrained, making it a useful test-bed for more challenging human behaviour interpretation. However, this is not to say the maritime environment is not without its challenges. Following from these motivations we can present a research objective which motivates and directs our work. Our objective is to investigate existing theory and algorithms with the capability to detect abnormal human behaviour in surveillance or maritime data, evaluate opportunities to improve the existing capability of such techniques, and propose and evaluate algorithms to better detect abnormal behaviour.

It is the three driving forces, listed above, that shape our goals in this body of work. The value of an anomaly detection system that is capable of detecting security relevant anomalies cannot be understated. Currently CCTV systems are

Figure 1.1: An example frame from the end goal of our research. We illustrate here the detection of 4 true positive abnormal behaviours. The 4 individuals highlighted by red bounding boxes are giving the purposeful impression of loitering and 2 are engaged in a suspicious bag drop. The 4 individuals (red bounding boxes) were classified as a social group (true positive).

primarily used as a forensic tool to determine the course of events after a situation. However the ability to monitor, in real time, the events and detect abnormal behaviour may allow for real time intervention for security events. Assets may be deployed accordingly, improving efficiency, and if timely enough preventing threats in first place. This body of work aims to take us a step closer towards this goal. This has been a goal of the computer vision community for a long time. It would be infeasible to realise this longterm goal within this body of work. As such we tackle more focussed goals as a way of making a contribution to the scientific community, hopefully bringing our knowledge of systems closer to this goal. We aspire to tackle and resolve particular aspects of the greater challenge by first studying existing methods, algorithms, and theory related to computer vision and surveillance to get a solid background understanding of how the challenge has been addressed to date. From this understanding of the work in the computer vision community we aim to identify gaps in theory that need to be explored and methods that need further scientific scrutiny. By doing this we will find ground for novel contributions towards the field.

1.2 Applications and Impact

1.2.1 Engineering Doctorate Program

This body of research was completed under the Engineering Doctorate program which spans 4 years of research with close ties to industry. The Engineering Doctorate program entails a preliminary taught year covering aspects of academia relevant to the research topic, as well as training in business through taught MBA courses. The goal of the Engineering Doctorate program is to build close working relation-

ships with industry partners, granting the student experience of both academic and industrial research. Furthermore it enhances the opportunity for the student to have real world impact with their research. We work closely with Roke Manor Research, a defence research company based in the south of Britain. For this reason there is a strong commercial drive behind our work. As a result of our research Roke Manor has developed their Accurate Target Tracking and Identification (ATTAIN) family of algorithms, which comprises a technology offering that has found its way into many customer funded projects and further research. Through ATTAIN much of the research and capability developed within this work has been exploited in real world applications. Our work has had significant industrial impact, however, due to the nature of defence research some of the applications of our research are restricted, and therefore cannot be detailed. We endeavour however to give details where possible.

1.2.2 Application of Research

The primary use of our research has been in a Maritime behaviour analysis project. This project aims to improve maritime situational awareness for large assets using a mixture of radar and Automatic Identification System (AIS) information. Our system is used to monitor the movement of other ships and small crafts to determine suspicious behaviour or threatening behaviour that should be brought to the attention of the operator. Our behaviour analysis method forms the long term behaviour analysis in combination with a faster short term motion abnormality detector which specialises in detecting acute motion anomalies. Our method applies the social model work to the maritime domain in order to detect groups of ships such as fishing fleets, convoys, and tugs. The importance of this step is that it is used to detect when a member of a group suddenly stops acting like the rest of the group, as this may be indicative of a ship hiding amongst legitimate behaviour. The intention is that this work can be used to increase the situational awareness and contribute towards the security of large military vessels, or port security. The project was funded through DSTL to the sum of nearly £250,000.

The human detection and tracking system we built for this thesis has been used as a means of detecting and tracking people through a small mounted head ups display for military use, similar to the kind that may be attached to a firearm. For this challenge we had to make the detection and tracking very lightweight to reduce the latency of the tracking to a minimum. The system would become inoperable with greater than 100 milliseconds latency. This application drove much of the optimisation of the people detection algorithms used in this body work.

The detection and tracking has been combined with re-identification algorithms developed in-house in order to carry out prolonged surveillance tasks in which the identity of people coming and going is of importance. However more details cannot be given about this project.

As there is a commercial drive behind our work we will take a practical approach and implement our theory and algorithms in this body of work. Implementation will allow for us to demonstrate the capability of our anomaly detection system upon representative data and evaluate the efficacy of our approaches. We are driven by the commercial background of our work to evaluate the receiver operator characteristic of our methods in particular, as well as demonstrate upon real world data cases of successful anomaly detection.

1.3 Approach

We must first discover what type of normal and abnormal behaviour exist in surveillance, what the characteristics and defining features of these behaviours are, how they are modelled in the state of the art, and where there is room for novelty to improve the current capabilities. We are ultimately motivated by the desire to contribute towards the ability of a computer to automatically interpret behaviour and detect intuitively abnormal behaviour from real world surveillance, with little to no human input. We review the current state of the art and historically relevant literature in the Background chapter of this thesis 2. Following from this chapter we find there to be a gap in the state of the art when using the social context of individuals in surveillance to enhance behaviour analysis. Furthermore we see the potential for using coarse head pose estimation to enhance the modelling of social connections and behaviour analysis. This finding drives much of our work as we seek to identify how head pose, and contextual information can be used to tackle hard anomaly detection cases and enhance the existing capability.

Our findings from the literature review lead us to the opinion that the computer vision community does not need another nuanced Hidden Markov Model (HMM) or machine learning algorithm. Of far more interest is the use of currently unexploited features. The crafting of a appropriate feature space has as much, often more, impact upon the accuracy of a machine learning system than the machine learning algorithm itself. There must be compatibility between the feature space and the machine learning algorithm. For example, a Nearest Neighbour algorithm requires a feature space which is a proper metric space; it must have proper non-negative and symmetric distance between any two points. Whereas a Neural Network requires that the feature space be expressible as a numeric vector. Once these basic restrictions are met there are many other considerations in order to maximise the performance of a system. Engineering a feature space which encodes the required information for the problem space, and is invariant to undesirable structure in the data, is of paramount importance. Ultimately the machine learning stage will learn the structure and dependencies of the data, and thus the desired structure, and the distinguishing classes, must be made as salient as possible. Our goal is to detect anomalies, and thus to make outliers as salient as possible. More recently the task of constructing an appropriate feature space has been automated using Deep Learning approaches [71]. Deep networks use regression to represent data at multiple levels of abstraction, allowing for complex models to be constructed. By masking some of the data or forcing the network to model the data with fewer nodes than the input, the network is trained to represent the data with a feature space that captures the principal components of the model.

We take the approach of adding to the catalogue of information that can be utilised by a machine learning algorithm for behaviour analysis. We use coarse head pose estimation as a means of determining the visual attention an individual has in their surroundings. Visual attention can be used to estimate social groups and characterise behaviours. On the theme of adding to the features that characterise behaviour we exploit contextual information in the form of social groupings and scene regions. We seek to demonstrate in this work that adding such contextual information into the representation of behaviour can enrich the understanding of behaviours and improve the detection of anomalies. Following from this background review of the field we can define our thesis as:

Feature rich, data driven anomaly detection algorithms can remove the

need for data intensive machine learning and expensive modelling techniques. By using contextual, motion, and head pose information we can separate heterogeneous behaviour clusters by increasing the interclass distance or reducing the intraclass distances, thus making outliers more salient. This allows for anomalies to be detected via the means of outlier detection.

Our hypothesis follows that we can offload the onus of the anomaly detection problem away from the algorithm that performs modelling and detection, and instead place it upon the representation of behaviour; on the features encoded, that are used represent behaviour. If the features model behaviour classes effectively then there will be little confusion between classes of behaviour; thus anomaly classes should be more salient by virtue of being less hidden by normal behaviour classes. We seek to answer this question with the following approach.

Feature extraction, detection, and tracking: We detect and track pedestrians within video surveillance data from publicly available sources. The scenes we use typically show semi-crowded public areas. In the feature exaction Chapter 3 we detect and track humans using our own system that draws from much of the state of the art work in these fields and is comprised from open-source code and our own. Our system uses the Histogram of Oriented Gradients (HOG) feature in a part based model to detect pedestrians in an image. A sliding window at multiple resolutions is used to partially overcome different pedestrian scales and locations, and multiple models are used to overcome 3D orientation variation. The detections are passed to a tracking stage which is modelled upon a short term motion tracker and bootstrapped object model detector. The purpose of the object detector is to reacquire tracks that are dropped due to occlusion or leaving the scene. From the tracks we can build pedestrian trajectories containing data about the position, speed, direction, and head pose at each frame of the track. The trajectories are passed to pre-analysis phase which derives additional information about social connections, scene dynamics, and performs feature signal processing. The sum of all the information is then passed to the behaviour analysis module.

Contextual information: The pre-analysis phase, Chapter 4 and 5 uses the motion and head pose information to derive additive information about the scene and the connections between those in the scene. We further progress the trend in context aware behaviour analysis by developing an unsupervised method of modelling scene dynamics. This method allows us to understand where in the scene different variations of behaviour are expected, such as fast moving or stationary motion. We additionally estimate social connections between pedestrians to further enhance our understanding of the behaviours in the scene. The social context information is designed around the idea of the social force being modelled as a spring between people; those individuals that are socially connected display closer proximity, mutual visual attention, and similar motion. The use of contextual information and exploitation of head pose information is a dominant theme in our work here. We use contextual information, head pose information, and motion information to take pedestrian detections all the way through to behaviour anomaly detection. The information derived in this stage is passed to the final behaviour analysis stage along with the motion and head pose information.

Anomaly detection: In this work we develop two behaviour modelling techniques, the second following from the findings of the first. The first method, in Chapter 4, verifies our hypothesis that contextual information, particularly social

and scene context, can improve behaviour analysis. The first method uses a simplistic nearest neighbour based approach to cluster behaviour. The second method refers to the Nearest Neighbour - Ranked Cluster Outliers (NN-RCO) procedure we propose and develop, Chapter 6. NN-RCO takes the behaviour metric further creating a metric space representation of behaviour which enhances the interclass distance using contextual information. We test this method on multiple datasets to verify the effectiveness and isolate the failure cases. We apply our methods to a different data domain, the maritime domain, in order to test the applicability to non-human behaviour. The maritime domain offers a different challenge in that tracking is trivial; each ship broadcasts regular Global Positioning System (GPS) coordinates and the motion is more constrained, however the abnormal behaviour is less intuitive, motion dependency is less salient, and behaviour can play out over days or even weeks. We wish to avoid developing an algorithm that is over-fitted to our particular data, or works only by exploiting some nuance of our collected data. Therefore our approach should be able to handle a broad spectrum of different behaviours and behavioural variation. We address this, in part, by using an adaptive approach; creating a system that defines normality relative to what it has seen before rather than using templates. However to test that our system is generically applicable we need to test on data that is characteristically different to the human surveillance data we targeted our system at. Testing upon the maritime domain addresses this.

In order to evaluate the feasibility and validity of our approach we implement the detection, tracking, social connection estimation, scene segmentation, and anomaly detection in Matlab and C++ and test upon several real world datasets. See Figure 1.1 for an example of one dataset. To ensure reliability, robustness and correctness of our system we base, where possible, our algorithms upon standard Matlab and C++ libraries, image processing toolbox and OpenCV.

1.4 Thesis Roadmap

Chapter 2 explores the current state of the art in several fields of study related to abnormal human behaviour detection. We make a judgement as to where more research is needed, and promising directions that the state of the art is making. Our work is concerned with the analysis of human behaviour. We thus evaluate literature in several disciplines relating to computer vision, particularly; human detection in video, target tracking in video, and anomaly detection.

In Chapter 3 we describe the process we use for extracting the features from imagery required in our behaviour analysis algorithm. We cover person detection, tracking, and head pose estimation, detailing the origin of the algorithm and implementation details of our particular use cases. We use the Deformable Part Based Model for person detection, which then feeds the TLD tracking component. Using the accurate head tracks from the previous tracking phase we extract head pose estimates using a supervised classifier. We bring together novel techniques from multiple sources into a single feature extraction process, additionally exploring improvements to the existing methods and optimisation. We make the following scientific contributions in this chapter:

- Integrating the intentional prior of head pose into pedestrian motion tracking, see Figure 3.4

- Validating that colour information improves the TLD tracking algorithm, and determining which colour space provides the greatest improvement, see Figure 3.3.4

- We propose and validate an alternative head pose classifier within the Benfold head pose estimation framework which has higher accuracy at increased computational cost, see Table 3.4.5

Chapter 4 proposes and investigates a system which leverages contextual information to improve the interpretation of behaviour and ultimately better find human behavioural anomalies in surveillance. We model human behaviour as a 2 part distribution containing a motion element which characterises the shape of the behaviour, and a context element which provides additional information separating subtle anomalies from the normal motion of behaviours. We use the data extracted in the previous chapter to demonstrate our method in 4 different surveillance scenes. We show that using an estimation of social connections in a scene (social context) and region classifications (scene context) we can improve behaviour anomaly detection. We evaluate our approach on real surveillance data and discuss the impact of the automatically generated contextual information upon automatic surveillance. The contributions made from this work are:

- A novel method of acquiring scene structure information in surveillance that compliments anomaly detection, see Figure 4.4

- The development of a novel social group classification algorithm using mutual information, see Figure 4.2

- The demonstration that social and scene contextual information can improve the detection of human behaviour anomalies; further validating the growing trend in automatic scene understanding, see Figure 4.6

In Chapter 5 we describe the development of an improved social and scene modelling method which builds upon our previous work, see Chapter 4 and state of the art techniques in Chapter 2. The aim of this work is to introduce the additional extractable feature of head pose, and the derived feature of visual attention into our social context and scene modelling work. Additionally we address the fundamentals of our previous context extraction algorithm, which proved the principle of using contextual information, to overcome some weaknesses and develop a more principled approach. We find that we can classify human social groups in surveillance at a higher accuracy with visual attention. Additionally an intuitive contextual model of the scene is developed which incorporates the head pose feature. We make the following contributions:

- The use of automatic visual attention estimation in social group classification system for surveillance, see Figure 5.1

- Evidence that social grouping is improved with the use of visual attention, see Figure, 5.4

- A method of deriving scene context information automatically, modelling the structure of the scene and comparative regional similarity, see section 5.4.1

In Chapter 6 we present our final human behaviour anomaly detection algorithm, NN-RCO, which builds upon all our previous research; incorporating in particular the social and scene context we previously developed. We detail the behaviour representation in section 6.2, how scene 6.3 and social 6.4 context information are used, and in section 6.5 we present the algorithm which detects abnormal human behaviour. In section 6.7 we demonstrate the feasibility and evaluate the proposed algorithm. We then provide a qualitative evaluation to the other state of the art techniques. This chapter comprises our primary contributions:

- The use of visual attention in a full human behaviour anomaly system, see section 6.5
- A novel anomaly detection system capable of including context information and simply integrating additional features such as visual attention
- A novel method for long term profiling of behaviour that elegantly handles tracking noise
- Evidence that subtle behaviours such as loitering and bag dropping have a visual attention element in their composition, see Figure 6.10

The purpose of Chapter 7 is to demonstrate the versatility and application of our algorithms in an alternative domain. This chapter serves to test the generic applicability of our approach. To test that our system is generically applicable we need to test on data that is characteristically different to the human surveillance data we targeted our system at. We test upon the maritime domain in order to assess this. The algorithm used is the same as human behaviour analysis algorithm from Chapter 6 with any adaptations outlined in section 7.3. The work we outline in this chapter derived from a real world application of our research and as such also demonstrates the impact of our research. Our primary source of data is the publicly broadcast AIS signal which presents GPS locations, speed, direction, and meta data for every ship within range. We capture the data with an aerial in house which gives range over Southampton and Portsmouth, in Britain. Our objectives for the maritime domain are the identification of suspicious behaviour in and around the background of legitimate traffic, apply algorithms capable of reducing operator workload, and to employ an algorithm capable of improving maritime situational awareness. The maritime work provides the following contributions:

- The demonstration of human behaviour analysis algorithms applied to the maritime domain
- Demonstration of the generalisation of our NN-RCO algorithm across data domains

We conclude in Chapter 8. We reiterate and evaluate how we have reached our aim for the this thesis. We bring together our work we presented in Chapters 3 through to Chapter 6. We first provide a detailed list of our contributions and conclusions from each section of the thesis 8.1. We follow this by outlining the theory and algorithms that were only partially investigated or remain to be investigated in section 8.2. We present a list of applications our research has had in section 8.3 and we finish with a final conclusion and remarks in section 8.4.

We next review the setting for our work by pulling together historically relevant and current edge theory, algorithms, and approaches relevant to the detection of

abnormal human behaviour in surveillance. We will use this review of literature to further define our objectives by adding specific goals for our research based upon our findings of the current state of the field.

Chapter 2

Related Work

In this chapter we assess the current state of the art in several fields of study related to abnormal human behaviour detection. We make judgment as to where more research is needed, and promising directions that the state of the art is making. Our work is concerned with the analysis of human behaviour. We thus evaluate literature in several disciplines relating to computer vision, particularly; human detection in video, target tracking in video, and anomaly detection.

The process pipeline required to get to human behaviour anomaly detection follows the sequence of human detection in images, tracking detections, additional feature extraction and derivation, and finally anomaly detection. We review literature in each of the disciplines, giving an overview of dominant methods and state of the art developments. We start with a review of the state of the art in human detection in section 2.1. We provide background to the human detection and basic concepts and theory that are needed. Section 2.2 introduces the task of object tracking in video, the problems associated with it, and state of the art techniques addressing the challenge. The merits and weaknesses of various techniques are explored. Section 2.3 presents the feature of coarse head pose, background and history of the feature, and state of the art methods for head pose estimation in video. This is followed by an exploration of anomaly detection theory and challenges in section 2.4. We review state of the art methods of detecting anomalies, anomaly metrics, and challenges in the field. Lastly we review contextual anomaly detection in section 2.5.

2.1 Human Detection

Dalal and Triggs [25] in their seminal work developed the Histogram of Oriented Gradients HOG approach towards detecting pedestrians in images. For a small local region the distribution of gradients can be encoded as a series of gradient histograms. The concatenation of all overlapping histograms over the entire object appearance forms the object descriptor. The system is capable of detecting pedestrians at varying size and appearance in a single frame. The HOG feature behind this approach works under the assertion that local object appearance can be described by the distribution and intensity of image gradients. The most common technique for extracting the image gradients is to convolve the 1D centred discrete derivative mask with the image. The mask consists of the kernels $[-1, 0, 1]$ and its transpose. Other filter masks can be used, such as the Sobel mask, consisting of the kernels and their transposes $[1, 2, 1]$ and $[0, -1, 1]$, however no benefit was found by Dalal and Triggs. Having defined the image orientation at each pixel the image is partitioned into cells. The cell's histogram is populated by the weighted votes of

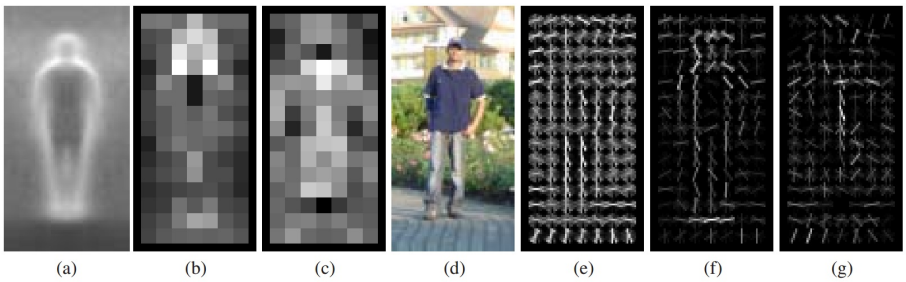


Figure 2.1: Image from the original work of Dalal and Triggs [25]. The figure illustrates the HOG training of a human detector. The HOG detector cues mainly on silhouette contours. (a) The average gradient image over the training examples. (b) Each pixel shows the maximum positive SVM weight in the block centred on the pixel. (c) Likewise for the negative SVM weights. (d) A test image. (e) Its computed R-HOG descriptor. (f,g) The R-HOG descriptor weighted by respectively the positive and the negative SVM weights.

each pixel within the cell based on the values found in the gradient computation, the weight of the vote is determined by the gradient magnitude. The standard HOG cell is rectangular, however radial cells are a frequent alternative. The gradients used extend to either $[0...\pi]$ or $[0...2\pi]$ radians. To overcome the impact of variation in illumination and contrast across the object appearance the cells are grouped into partially overlapping blocks of cells and then normalised across spatially connected blocks. Dalal and Triggs found the optimal configuration to be blocks consisting of 3x3 cells, each cell consisting of 6x6 pixels, with 9 histogram channels covering 0 to π gradient orientations. Additionally it was found that applying a Gaussian weighting to the votes of pixels in each block, lowering the voting power of pixels nearer the edges of blocks, improved performance. Cell normalisation across a block is achieved using L2-norm, defined as:

$$L2 - norm = \frac{\mathbf{v}}{\sqrt{||\mathbf{v}||_2^2 + e^2}} \tag{2.1}$$

Where e is a small constant and $\mathbf{v}$ is the unnormalised descriptor vector. The final step in the human detection process is to apply a linear Support Vector Machine (SVM) classifier [25] to query image patches. Once trained on images of humans the SVM can make the binary decision as to whether the query image contains an image of a human or not.

Dalal and Triggs train a HOG feature vector upon 509 instances of standing pedestrians and then a SVM is used to classify the subsequent 200 test images. The approach scans an image at multiple scales. This HOG-based approach was extended to deformable part-based object detection, developed by Felzenszwalb [30]. The traditional histogram of oriented gradients-based pedestrian detection outlined in Dalal and Triggs [25] uses a single HOG model to assess how similar the feature vector is to human appearance. Felzenszwalb extends this to using multiple parts to the model. Model parts are located in the image and the detection confidence is scored by the closeness of fit to the model and the deviation from expected distance to each other. This method allows the model to deform at a cost. The method is slower yet far more accurate. Junjie progresses the DPM method by speeding up the bottleneck in the procedure [87]. The 2D correlation is constrained to a low rank combination of 1D correlations. Instead of explicitly calculating all part scores, a

neighbourhood aware approach aggressively prunes parts with an understanding of dependence between neighbouring regions. Additionally look up tables are used to replace the more expensive computation of gradient orientations. The method is up to four times faster than standard DPM with similar accuracy.

There has been a recent surge in the use of Convolutional Neural Networks (CNN) for object detection and classification tasks. Not only do they provide a jump in accuracy, but they provide additional capability to select the relevant features from the data for the detection task. Such approaches use an almost brute force approach to classification by searching the feature space for useful features to perform the classification task. Girshick [33] uses a convolutional neural network in part of a three module approach to person detection. The first section returns candidate detections for the detector. The second module is a large CNN that extracts the feature vector from each candidate detection and passes the vector to the third module. The third module is a linear SVM. This approach exceeds existing state of the art approaches, achieving a mean average precision 30% greater than the previous best results. The main novelty of this approach is the use of applying a high capacity CNN to detection proposals in order to segment and classify pedestrians. Additionally, context can be exploited as a prior to enhance object detection as demonstrated in the work by Mottaghi et al [65]. Pixels are labelled with a semantic category, and a deformable part-based model which exploits the local context around a potential detection and global context in the image is used to detect.

Many methods circumvent the need for human detection by using a non-agent behaviour representation which does not pertain to the individual [39], [61], [76], [5]. Such methods typically use optical flow to estimate motion in the image. Optical flow measures the relative pixel movement of pixels to the observer. For a static camera this measures the movement in the scene, and for a moving camera there is ambiguity of scene motion and camera motion which can be resolved using a calculation of parallax from a moving camera. Typically optical flow is used to measure crowd dynamics [64] and other aggregate motion features [75].

2.2 Human Tracking

Tracking divides into two categories of particular interest for surveillance; point tracking, and appearance tracking [19]. Point tracking encompasses methods which track an object as a cluster of independent or jointly distributed points in the image plane. Appearance tracking, also called kernel tracking, represents the object being tracked with a feature vector or area on a manifold to allow association between candidate locations in subsequent images. Silhouette tracking works by tracking the transformation of the object perimeter in subsequent frames. Point tracking methods are particularly applicable to Wide Area Motion Imagery (WAMI) where an object maybe represented by only a few points due to the distance from which imagery is taken. In the work of Jiejie et al [89] pedestrians are detected by their motion. This approach is necessitated as the target size is only a few pixels in size. Multi scale intrinsic motions structure is extracted for each pedestrian. The target track consists of a ground plane coordinate and velocity vector at each frame; the intrinsic motion structure is calculated by taking the principle axis from a tensor clustering of the set of motion vectors for a track. The index of the maximal eigenvalue difference defines the dimensionality of the motion and the difference defines the saliency. The clustering neighbour size gives the scale at which the motion is ex-

tracted. Pedestrian detection is then treated as a binary classification problem using Adaboost to train the classifier. Pedestrian tracking is achieved in [14] by tracking points that are associated between frames by their HOG descriptor. The object motion can then be resolved from the trajectories of the tracked points. To handle occlusion the direction, speed, and displacement of the HOG point trajectories are compared with those of objects in previous frames to determine the bounding box split for the occluded object.

Tracking using appearance models largely divides into two approaches; tracking by motion, and tracking by detection. However, an exception to this is the work by Kalal [42]. Kalal uses a hybrid of tracking by motion and tracking by appearance to create stable unconstrained object tracks with re-detection. An appearance model is bootstrapped from the first frame by positive and negative learning to define a region on an appearance manifold that represents the object appearance. Short term tracking is achieved using Lucas Kanade (LK) optical flow tracking. LK tracking is carried out in parallel with the bootstrapped object detection and the higher confidence response is used as the object location in the next frame. The strength of the approach lies in the ability for one tracking thread to compensate for failure in the other. A recent tracking by detection approach was developed by Benfold [11] in which pedestrians are detected in a two strand process running in parallel. One side returns detections asynchronously using a HOG pedestrian detector similar to the Dala and Triggs person detector [25] and provides a small amount of motion based tracking both forwards and backwards from the detection. The second part classifies detection as false detections or person and associates the true detections to track identities based upon size, velocity, and positionally similarly using the Monte Carlo Markov Chain approach. The approach can handle a variable amount of time between detection frames thus adapting workload to maintain video rate processing. In [18] object appearance parameters are learnt offline for each tracking context. Features are selected via learnt weights in an offline step which segments the different tracking components; such as 2D and 3D displacement, size, colour, HOG, colour covariance, and dominant colour. Detected objects are then associated across time based upon this appearance.

2.3 Head Pose Estimation

We utilise a more specialised feature, head pose, in our research. The head pose of an individual can be used to make an estimate of the individual's visual attention. The extraction of head pose in surveillance is distinct from gaze localisation which determines the focus of the eyes in a constrained environment; typically with a high resolution camera placed in front of the subject. Instead, our goal is to extract head orientation from low resolution (ranging from 10 pixels in height) head images in surveillance, which is fraught with occlusion, lighting variation, and non-parametric appearance. The extraction of head pose from surveillance is a relatively new area of study. Gourier uses Grey-level normalized face imagettes as features for a linear auto-associative memory, where a single memory is computed for each head pose using the Widrow-Hoff learning rule [34]. Robertson [74] built parametric models of skin, hair, and background from hand labelled examples. A decision tree is used for head orientation classification into 8 difference possible angles, where the decisions are the hair, skin, or background class that specific pixels fall into. This work was progressed by Benfold [13] enhancing the head pose estimation to be colour invariant and un-supervised. This method used pixel triplet comparisons as the

Figure 2.2: Illustration of head pose classification in the surveillance. Head pose is indicated by a field of view cone extending from the head bounding box.

binary decisions in a forest of decision trees. The decision trees are trained using a weakly supervised method where the direction of travel is assumed to be an indicator of head orientation. Through subsequent iterations of weakly supervised training the classifiers converge. A more recent trend in head pose estimation is to build a joint distribution over both head and body pose as there is a clear dependency between the two [49, 20]. In the work of Krahnstoever body and head pose is estimated independently of the direction estimate using a combination of sequential Monte Carlo Filtering and Monte Carlo Markov Chain sampling which encodes the propagation of horizontal head angle to body pose [49]. In a similar approach Chen and Odobez estimate both head and body orientation modelling the dependency as a joint model adaptation problem. Head and body pose are estimated by appearance classifiers which are learnt in a weakly supervised fashion using motion as an initial estimate [20].

2.4 Anomaly Detection

The concept of anomaly detection is potentially vague as it does not necessitate any particular detection, classification, or statistical approach. Thus we must first carefully define the meaning for anomaly detection for the context of our work. There are many definitions of 'anomalies' that can be found in the literature. Tan et al. [79] define anomaly detection as the task of detecting observations whose characteristics are significantly different from the rest of the data. Loy [60] defines an anomaly as an event which has a low statistical representation in the training data. Chandola et al. [17] defines an anomaly as a pattern in the data that does not conform to a well defined notion of normal behaviour. The goal of anomaly detection is thus to classify outliers in a given dataset. Outliers are descriptive of variations due to noise, deviations, exceptions in the data, and contradictory behaviour. To pin down the definition of an anomaly we take the definition given by Loy [60], defining an anomaly as an event which has a low statistical representation in the data. Thus anomalies are defined in terms of not being represented by a function modelling the majority of the data, given a dataset. This implies a process which models normality in the dataset, which by contrast anomalies can be detected. A simple introduction to the concept can be found in Figure 2.3 in which normal events are modelled

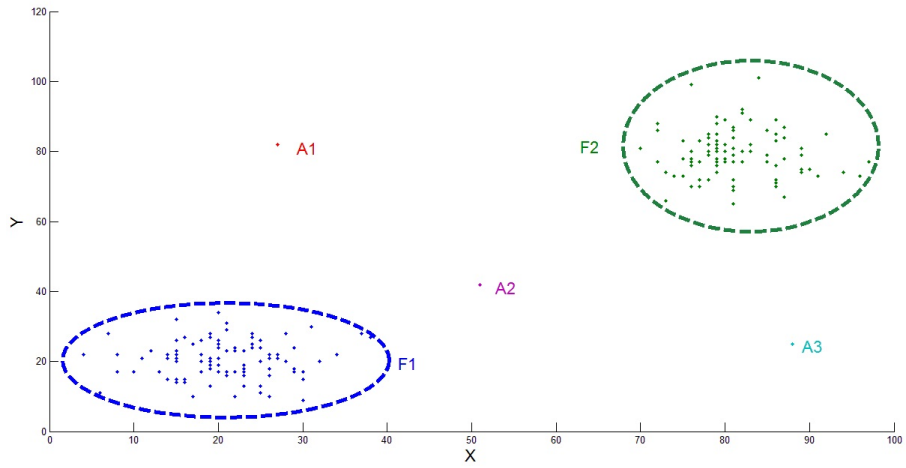


Figure 2.3: Simple example of anomalies in an arbitrary space. Events are modelled (spatially) by function $F1$ and $F2$. Anomalies are assigned as $A1, A2, A3$.

(spatially) by function $F1$ and $F2$. Anomalies are assigned as $A1, A2, A3$. The key to an effective anomaly detection system is modelling the data such that outliers are separated from the mass of normal data. A effective representation of events will decrease intraclass distance, and increase interclass distance, such that events are closer clustered to similar events, and further clustered from dissimilar ones. Anomalous events, distinct from all others and poorly represented in the training data, will thus be easier classified as outliers. This concept is one that drives much of our work in later chapters.

2.4.1 Challenges in anomaly detection

We can imagine all events in a video sequence as being represented in an arbitrary 'event space' where each event has a particular position or is a distribution within the event space. Dimensions in the event space would perhaps represent observable features or derived information characterising the event. Anomaly detection is concerned with defining a function which represents regions in the event space which encompasses normal events, and from which we can determine the degree to which any event is an outlier. This can be visualised similar to Figure 2.3. The task of anomaly detection becomes more challenging when noise is injected into the system. Spurious noise patterns may warp the appearance of a legitimate behaviour, giving the impression of an anomaly, or indeed noise may mask the occurrence of an anomaly. A particular difficulty facing anomaly detection is that of defining a function which specifies a region on the event space which captures all of the legitimate variation in normal behaviour. Particularly in natural behaviour sequences such as human surveillance the scope of possible normal behaviour and the variation in representation of these normal behaviours can't be captured by any simplistic function definition. Furthermore the realm of normal behaviours may be changing over time. We may have a dynamic scene in which the regions for certain behaviours or events may migrate in space, change characteristics gradually over time or even suddenly, or the interpretation of the same event may change from normal to abnormal. Consequently, the nature of anomalies may change over time. For agent driven behaviour it is even possible for an agent to actively adapt and attempt to mask abnormal behaviour. The rarity of training data is often a hindrance in anomaly

detection. With no prior definition of an anomaly, the characteristics of anomalies must be derived from the statistical dissimilarity to the majority of events, which necessitates a quantity of training data from which significant trends can be extracted. Given these numerous confounding factors solutions must tackle most or all of these problems either implicitly or explicitly. There are several practical issues which must be additionally overcome when implementing an anomaly detection system, particularity in industry; computational efficiency, subjectivity in acceptance of anomalies, feedback and justifying anomalies, and achieving an acceptable false positive rate.

2.4.2 Types of Anomaly

It is important to consider the different types of anomalies that may be present. We tailor the following examples for the human surveillance domain, although the list is still fairly generic. We largely follow the work of Chandola et al. [17] in the following.

Point anomalies: Point anomalies are the simplest form of anomaly, consisting of a single instance of data that resides in a region of the event space. Typically this region would be defined by being distant to any examples of normal data points. Figure, 2.3 is an example of such point anomalies. The exact significance of the distance of a particular anomaly is determined by the distance metric of the event space.

Extended anomalies: An extended anomaly is a collection of information relating to a single event in the data that extends over a region within the event space. The information may pertain to a collection of instances of a single agent extended over time, a distribution over an area representing uncertainty, or a collection of related features representing the state of an event. A typical example of such an anomaly could the trajectory of an abnormal individual in surveillance represented as a distribution over possible behaviour states.

Collective Anomalies: Collective anomalies consist of a cluster of intrinsically linked events, which may on their own not be anomalies, but when considered in relation reveal the existence of an anomalous event. An example of which may be an environment on which approaching entering and leaving a car are all instances of normal events. However the collection of approaching a car and then approaching another and then another is abnormal. In such a case the probability of transitioning between normal events must be considered to reveal the anomaly. Alternatively, it may be normal for any individual in a scene to look in any particular direction, however when everyone in the scene all look in a particular direction it reveals the existence of an abnormal event perhaps even out of the coverage of the camera. In such a case the co-occurrence of events must be modelled.

2.4.3 Classification

It is not necessary to classify every event into either 'abnormal' or 'normal' as often the events are left with an outlier score; a continuous variable representing the degree of separation to the set of normal behaviour. The results can then be presented as a ranked watchlist or visualised as a heat map of anomaly probabilities. When an anomaly classification is desired there are three dominant methods of training the system; supervised, semi-supervised, and unsupervised training.

Supervised: Supervised methods use existing class labels to train a classifier.

Typically such method use hand labelled normal and abnormal events. However there are a number of drawbacks to this approach. Firstly, the number of instances of normal events normally far outweigh the instances of abnormal. Due to this imbalance it can be hard to obtain a representative set of examples for abnormal events. The under-represented model of abnormal events may skew classification results if insufficient information is known about the characteristics of the event. Furthermore, there may not be training instances of all abnormal or normal events, requiring the system to extrapolate potentially incorrectly or miss events altogether.

Unsupervised: Unsupervised classification techniques do not rely upon labelled training data, but instead impose class labels automatically. Unsupervised anomaly detection methods are based upon the assumption that the frequency of normal behaviour is far greater than that of abnormal cases, and thus abnormal events can be identified by their rarity or lack of statistical representation. Furthermore, abnormal events must be distinct from normal events by some feature or characteristic visible to the system. Typically such systems must make allowance for the fact that the training data may contain instances of abnormal events. Over-fitting is a danger for most unsupervised methods.

Semi-supervised: Semi supervised approaches either use a mixture of labelled and unlabelled data or use weakly labelled data [10]. Often some of either the normal or the abnormal events are labelled instructing the system how to segment the data into normal and abnormal. The weakly supervised methods use a large amount of automatically labelled data with known error rate which does not overwhelm the training data. Such methods are applicable when class labels can be estimated and hand labelling is prohibitively expensive.

The output of an anomaly detection system depends on the nature of the events being classified. In human agent surveillance the output is typically displayed to the operator as bounding boxes around those performing abnormal behaviour. However it may be informative to display the confidence of the anomaly in which case a colour intensity often represents confidence. This is particularly the case when there is no absolute anomaly classification and all agents hold an anomaly score.

2.4.4 Anomaly Metrics

The task of classifying an anomaly can be considered that of assigning a class label; there may be multiple anomaly classes and multiple normal event classes. There are a common set of metrics used to measure class similarity in anomaly detection; model-based classifiers, Nearest Neighbour, clustering techniques, statistical classification, spectral methods, and information theory.

Model Based Classifiers: Model-based classifiers use a descriptive model to express the difference between two classes and a predictive model estimates the class label. In supervised methods the descriptive model learns from labelled training data. The predictive model is then used to assign a class label to new data based upon the descriptive model output. As an example of model-based classifiers, network anomaly detection methods, are based upon network traffic models [3]. Hajji presents a descriptive Gaussian mixture model, using a stochastic approximation of the Expectation-Maximization algorithm to obtain estimates of the model parameters [35]. Prediction of abnormal events is achieved via a decision threshold. A widely used method is that of Neural Networks. Stefano trains a multi-class neural network using training data of all normal classes [78]. Query data is then classified by feeding it through the neural network. Query data that was classified as one

of the training data classes was considered normal, without the need for normal class labels. Weiming et al. [38] conducts anomaly detection via activity understanding using a fuzzy self-organising Neural Network. Weiming presents a method for learning patterns of object activities in image sequences. Activity patterns are modelled using unsupervised learning of motion trajectories. Unlike many neural network-based methods they use the whole trajectory of a target as an input to the network, making the network structure much simpler. Another technique that has spawned many variants is the Bayesian Network approach to anomaly detection. A Bayesian Network is a simple technique for constructing a classifier model that assigns class labels to behaviours represented as vectors of feature values, where the class labels are drawn from some finite set. In the case of univariate, conditionally invariant data, a Naive Bayesian classifier can be applied [68] which assumes the values of features are conditionally independent. An alternative method for classifying between two classes is that of Support vector Machines. A support vector machine learns a hyperplane which segments two or more classes in, typically, a multi dimensional feature space. Test data is classified as either class depending on which side of the hyperplane the data falls. Wenjie et al. compares the performance of Robust Support Vector Machines (RSVM) with that of conventional support vector machines in separating normal usage profiles from intrusive profiles of computer programs [86]. RSVM address the problem of over-fitting which can occur due to noise in the training data set with an averaging technique which makes the decision surface smoother.

Nearest Neighbour: Nearest Neighbour search is an optimisation problem focussed upon minimising a distance metric over all possible pairings of data points to a query point. Specifically, for data Y there is a data point Y_m which is closest to Y_n:

$$NN(Y_n) = \{Y_m \in Y | \forall Y_p \in Y : \Delta(Y_n, Y_m) \leq \Delta(Y_n, Y_p)\} \tag{2.2}$$

Nearest neighbour anomaly detection methods work upon the assumption that normal data is clustered and as such similarity is a suitable metric to determine outlier distance. The crucial aspect of a nearest neighbour algorithm is the definition of a distance function between data points. For human behaviour anomaly detection the behaviour is represented in a metric space where the distance between any two behaviours is well defined. The simplest solution to solving the Nearest Neighbour search is to compute the distance from the query point to every other point in the database, or linear search. This naive approach has a running time of $O(Nd)$ where N is the number of data points to search over and d is the dimensionality of data. Several methods exist for optimising the search algorithm; Space partitioning, Locality sensitive hashing, Vector Quantisation, and Greedy walks. There are two primary classes of nearest neighbour-based anomaly detection. The first uses the distance between a data point and its k-nearest neighbours as the outlier metric. The second class uses the relative density in a neighbourhood as a the outlier metric, which can be imagined as the radius of a hypersphere encapsulating the k nearest neighbours divided by k.

Clustering Based Techniques: Clustering techniques find groups of similar data, effectively automatically labelling the data into classes. There are three main families of clustering techniques. The first makes the assertion that normal belongs to a cluster, and abnormal data will be an outlier to *all* clusters. Techniques such as DBSCAN [29], ROCK [81] and SNN [58] fall into this category. The second group of clustering techniques make the assertion that normal data instances fall

close to the centroid of their nearest cluster, whilst abnormal data instances end up far away from the local cluster centroid. This method requires an initial training stage clustering the training data. Following from training the distance from each data element in the test data and the closest cluster centroid is calculated and used an an outlier score. Methods such as Self-Organising Maps [48], K-Means Clustering [16], and Expectation Maximisation Clustering [26] fall into this category. The third method of anomaly detection via clustering asserts that normal data clusters into large clusters whilst abnormal data clusters into small or sparse clusters. An example of such a technique is Cluster Based Local Outlier Factor [36]. These clustering techniques share properties with that of Nearest Neighbouring Anomaly detection. The distance metric in the metric space that events are represented is the determining factor in how well the method classifies anomalies. Anomalies are similarly defined by the degree to which they are an outlier to the main distribution of the data. The main difference lies in clustering techniques requirement in the appearance of neatly defined clusters. Nearest Neighbour approaches do not require clusters to form; the data can be sparse or uniform. Clustering techniques do not, in general, require data to be labelled as they take advantage of the natural structure of the data. However clustering techniques can be computationally expensive, typically falling into $O(n^2)$ complexity.

Statistical Techniques: Statistical techniques train a model on example data to classify new instances of data into a trained class. A query datum that is generated from the same stochastic process as the training data is expected to fit the statistical model well, and data not from the same process will not fit, and will thus be classified as an anomaly. Statistical methods of anomaly detection fall into the categories of either parametric or non-parametric. A parametric technique assumes knowledge of the underlying distribution of the data, and as such can make inferences from a small amount of data based upon known parameters. Unlike parametric statistics, non-parametric statistics do not make assumptions about the data distribution. Non-parametric statistical techniques use less information in their calculation. For example, a parametric correlation uses information about the mean and deviation from the mean while a non-parametric correlation will use only the ordinal position of pairs of scores. Parametric methods often assume an underlying common distribution, such as; Gaussian, Poisson, or Binomial. Examples of which are Gaussian Mixture Models, or Regression Models. Non-Parametric methods, making no prior assumptions about the data distribution, include methods such as Histograms, Kernel Density Estimation, Non-parametric Regression, Data Envelope Analysis, and K-Nearest Neighbour. The computational complexity of the approach depends on the statistical model.

Spectral Techniques: Spectral techniques are based upon the principle of dimensionality reduction. The motivation is to reduce complex data to the principle information, at which point abnormal instances of data may be readily identified. A commonly used method is that of Principal Component Analysis (PCA). This technique reduces the correlated dimensions of a dataset while preserving the variation. The process reveals a number of statistically uncorrelated principle components ordered by variance. By removing redundant and possibly misleading information from the dataset new data instances can be evaluated by how well they fit the principle components. Those data instances that do not fit can be regarded as anomalies. Similar to PCA is Compact Matrix Decomposition (CMD) [82]. CMD is used to compute sparse low rank approximations for revealing latent/hidden variables and associated patterns from high dimensional data. CMD reduces both the computa-

tional cost and the space requirements over existing decomposition methods.

Information theoretic: Information theoretic techniques use measures of information entropy, conditional entropy, relative entropy, minimum description length, and Kolomologrov complexity. Such methods work upon the principle of Occam's razor; that a simpler solution is more accurate than a more complex one. Anomalies that do not fit the common characteristics of the data increase the complexity of the data, which requires a more costly solution to express the dataset. Minimum description length techniques in particular follows the ideas of data compression to find the optimal class labels for all data entries to minimise the information required to describe the data set. This approach is similar to clustering techniques however the system not only considers the cost of grouping any two instances of data but also the cost of not grouping data, as not grouping similar data will increase the complexity of the dataset description. Eberlea et al. [27] use minimum description length to model the normative pattern in large datasets which can then be passed to an alternative anomaly classification algorithm. Alternatively entropy can be used as a measure to isolate anomalies in a dataset X, where each datum belongs to a class $x \in C_x$. We can define the entropy of the set relative to the $|C_x|$ classification as:

$$H(X) = \sum_{x \in C_x} P(x) \log \frac{1}{P(x)} \tag{2.3}$$

Where $p(x)$ is the probability of x in X. We can interpret the entropy of set X as the number of bits required to encode the classification of the data. Entropy is smaller when the data conforms to a small distribution over possible classes, and larger when there is a greater disparity. If all data belong to a single class then the entropy is 0; it takes 0 bits to encode the dataset as there is only 1 possible solution. If all datum are evenly distributed over all classes then it takes $log|C_x|$ bits to encode the set. Entropy can be used in anomaly detection as a measure of regularity in a dataset. The lower the entropy the fewer the number of different classes in the data. Highly regular data contains redundancy which means that future events are more predictable as low entropy suggests they are likely to be repeats of current classes. An abnormal event in a data stream can be detected by deviation of established complexity. Conditional entropy provides a method of better measuring the temporal or sequential nature of data to better classify temporally dependant anomalies:

$$H(X|Y) = \sum_{x,y \in C_x, C_y} P(x,y) \log \frac{1}{P(x|y)} \tag{2.4}$$

Relative entropy gives a measure of how well a distribution of training data matches a distribution over test data. It is in effect a batch process calculation of conditional entropy:

$$R(p|q) = \sum_{x \in C_x} p(x) \log \frac{p(x)}{q(x)} \tag{2.5}$$

A smaller relative entropy suggests that the test data closely fits the training data [57] predicting fewer anomalies in the test data.

2.5 Context Aware Anomaly Detection

The metrics for anomaly detection provide a means of classifying and clustering observations of behavioural events. However, this is only part of the anomaly detection task. We must also address the way in which we express behavioural events. How we encode the characteristics and observable information pertaining to an event opens the scope for how we can classify the event and similar events. A growing trend in human behaviour anomaly detection is that of context aware anomaly detection. With context we can augment our representation to incorporate additional information which changes, further refines, and adds to our understanding of the observable information. Context information can be manually sourced, or more interestingly automatically derived from the data. Particularly in dynamic scenes, contextual information such as changes in behaviour spatially or temporally can be particularly informative. We focus upon social and scene region contextual knowledge as a means of improving the detection of subtle behavioural anomalies in our own work. The scene regions provides an understanding of portions of the scene in which we would expect normal behaviours to be different from other areas [63]. Previous approaches such as Li et al. develop a scene segmentation method which divides the scene into regions based upon behavioural dissimilarity [59]. Similarly, Loy segments a scene into spatial regions of similar behaviour by virtue of behaviour correlation [60]. This work introduces a second line of contextual scene knowledge: temporal state. This contextual information is particularly apt for the traffic junction, in which behaviour is clearly temporally segmented in short time intervals. However, it is far less applicable to many human surveillance environments where the periodicity of behaviour is far less structured, if at all. Wang et al. uses a Dual Hierarchical Dirichlet Process to cluster behaviours spatially, learning both observation and trajectory clusters simultaneously [85].

The second source of contextual information we use is social context. Social Context grants the ability to learn the distinction between normal behaviour for groups and individuals independently. The social model provides an additional benefit; it ensures that the behaviour of each individual is analysed in reference to people external to the same social group. Thus a homogeneous group of individuals all acting abnormally can't be self-justifying. Furthermore social information enables us to create likelihood dependencies between individuals in a social group. Thus if one individual in a group is behaving abnormally the expectation of other group members behaving abnormally goes up. The estimation of social groups in surveillance has a focused primarily on motion features. To estimate social groupings Ge et al. uses a proximity and velocity metric to associate individuals into pairs, iteratively adding additional individuals to groups using the Hausdorff distance as a measure of closeness [32]. Yu et al. implements a graph cuts-based system which uses the feature of proximity alone [88]. However modelling social groups by positional information alone is perilously primitive and prone to finding false social connections when individuals are within close proximity due to external influences such as queuing. Oliver et al. uses a Coupled HMM to construct a-priori models of group events such as Follow-reach-walk together, or Approach-meet-go separately [66]. Certain actions are declared group activities and thus groups can be constructed from individuals via mutual engagement in a grouping action. However, a more recent development in automatic social grouping seeks to model social interaction using the visual interest of the tracked individuals. The use of an individual's visual attention is significant as it uses a rich feature which indicates the intention of the individual. Robertson

and Reid utilise head pose direction in order to determine whether individuals are within each other's field of view [74]. Farenzena et al use an estimation of the visual focus of attention of a person as a cue to indicate social interaction [9]. Head pose is quantized into 4 different locations at each frame, and a predefined set of spatial and visual criteria determines if the conditions for a social interaction are met at each time step. A social exchange is then defined as lasting a given duration (10 seconds). In our work we bring together the motion-based social paradigm with the benefit of visual information as it is demonstrated by [74] [9].

2.6 Human Behaviour and Visual Attention

We previously reviewed anomaly detection based upon motion features. We now extend this review to include work using visual attention. Visual attention in this context covers the use of head pose information (also called gaze direction) and more complex derived estimates of attention. Visual attention does not have a long history in computer vision, less still when applied to behaviour analysis. Stiefelhagen [80] tracks attention in a meeting and cues camera motion to automatically focus upon the primary speaker based upon the focus of attention estimation. Focus of attention is based upon head pose and eye tracking simultaneously. In a complex scenario with four speakers the correct focus of attention can be estimated from head pose alone with an accuracy of 87%; lending further credibility to the idea of estimating visual attention from head pose alone in surveillance. Odobez [6] furthers the preceding work by generation of a social meeting model enhanced by head pose direction estimation combined with motion and proximity information, and contextual information such as whether or not targets were engaged in conversation. Farenzena et al. [9] uses head pose direction to estimate a 3D region of visual attention in a scene which is subsequently used to classify social interaction. Similarly an early example of head pose being used to supplement an representation of behaviour is that of Robertson [74]. Robertson suggests a causal reasoning process based on a set of qualitative facts drawn from observations of action, behaviour, and head pose estimates. He generates qualitative text descriptions of a scenario, such as a tennis match. The text description are obtained automatically from the action and behaviour recognition stages of the system. Particularly in the tennis example, where head pose is significant, the system is shown to reason causally about sequence of actions observed. The system developed is rule-based and as such action classification is restricted to a sequence taken from the set of prior known actions. Sequences of actions are combined through the use of a Hidden Markov Model into behaviours. Benfold [10] maps visual attention in multiple scenes using a novel implementation of head pose estimation. Initially an accumulated map of visual attention is built from the analysis of the Oxford Town Centre dataset [12], which covers a busy town centre street with up to thirty pedestrians visible at a time. Regions of high visual attention were identified on shop fronts in particular. Further experiments were carried out by artificially drawing the attention of pedestrians to a particular point by attaching a light to the wall. Attention maps were then constructed with the light off and with the light on. The difference between the two attention maps clearly indicated the presence of the light source. Further research identified transient objects of interest by tracking visual attention of multiple targets and finding the intersection of head pose direction estimates. An example being the highlighting of a moving vehicle.

2.7 Conclusion

Two distinct positions dominate human behaviour anomaly detection in surveillance. The first defines behaviour as an agent activity and builds a human centric behaviour description. Humans are detected and tracked as discrete entities and behaviours revolve around the trajectories and interactions of the agents. The second, non-human centric behaviour, seeks to define anomalies as patches of motion in the image stream. With non-human behaviour representation a typically positionally fixed model learns normal motion patterns that defines a region; anomalies are described by changes in motion or appearance of the foreground. Both approaches have their merits; non-human centric behaviour has advantages in crowded or highly occluded scenes, or scenes with a high range of classes of agents. However, without including the notion of the individual in the representation the descriptive capability does not align well with human intuition. Human centric approaches have the advantage of encoding more information about the interaction of the agents responsible for behaviour, longer term profiling can be achieved, and tracking of humans can lead to further features such as head pose and contextual features being derived.

The dominant trend for Human detection in video is that of Histogram of Oriented Gradients-based detection. We find that the current best adaptation are part-based models using HOG features. Deep learning techniques show much promise, however. Most human centric methods follow the same pipeline; tracking typically follows a detection phase, then tracks are passed to an anomaly classification technique. There is a great variety of tracking algorithms to suit various nuances of environments and objectives. In terms of overcoming target loss and occlusion in a semi-crowded surveillance scene, re-identification of targets is paramount to a stable tracking system. Online learning trackers have proven robustness to occlusion when given ample time to build an object classifier. The TLD tracker shows much evidence of effectiveness at overcoming tracking occlusion and successfully re-identifying dropped targets. Our evaluation of anomaly detection procedures suggests model-based classifiers are dominant in data and information anomaly detection, however nearest neighbour and clustering approaches are dominant in human anomaly detection, perhaps in part due to not requiring labelled class data, self organisation, and implicit modelling of the behaviour types. The feature of head pose has received interest recently, particularly in the methods of extraction from video. More robust colour invariant techniques have emerged permitting the use of coarse head pose estimation in low resolution surveillance in anomaly detection.

There exists a gap in the state of the art when considering the implementation and analysis of automatic contextual information in human surveillance. In particular there is an opportunity to enhance and evaluate human behaviour anomaly detection using social modelling and scene understanding in surveillance. We also observe that many context aware techniques do not derive *additive* information, but instead re-factor existing information. Furthermore we are yet to see a method which implements and utilises contextual information about social connections to better classify abnormal behaviour in human surveillance. Additionally, although techniques do exists, there is scope to demonstrate the efficacy of scene modelling in human behaviour anomaly detection. The recent advances in head pose extraction have opened the way for the use of head pose information in social modelling, scene modelling, and behaviour analysis. There is a definite gap in the state of the art with regard to evaluation of the effectiveness of head pose information, and derived visual attention estimation in human behaviour anomaly detection.

In this chapter we have identified the current common and effective properties of human behaviour anomaly detection algorithms. We reviewed state of the art and historically relevant anomaly detection algorithms applicable to human behaviour anomaly detection and identified where there is a gap in current understanding and opportunity for improvement. In light of the findings from this chapter we can propose the following research objectives:

- Objective 1 Propose algorithms to deliver additive social context information into an anomaly detection system
- Objective 2: Propose algorithms to deliver additive scene context information into an anomaly detection system
- Objective 3: Propose a novel algorithm for determining human behaviour anomalies which integrates contextual information into the analysis
- Objective 4: Demonstrate the entire pipeline of our proposed algorithm upon real world surveillance data
- Objective 5: Demonstrate the feasibility and quantify the effectiveness of contextual information in human behaviour anomaly detection on real world data
- Objective 6: Implement head pose estimation and utilise the information in our contextual work and behaviour analysis
- Objective 7: Evaluate our proposed algorithm upon real world surveillance data, demonstrate the efficacy of our approach and assess our algorithms in light of other state of the art approaches
- Objective 8: Evaluate and quantify the effectiveness of head pose information in contextual information sources
- Objective 9: Evaluate and quantify the effectiveness of head pose information in human behaviour anomaly detection

Chapter 3

Human Motion Feature Extraction

In this chapter we describe the process for extracting the features from imagery required in our behaviour analysis algorithm. We cover person detection, tracking, and head pose estimation, detailing the origin of the algorithm and implementation details of our particular use cases. We use the accurate Deformable Part Based Model for person detection, which then feeds the Tracking Learning Detection tracking component. Using the accurate head tracks from the tracking phase we extract head pose estimates using a supervised classifier. We bring together novel techniques from multiple sources into a single feature extraction process, additionally exploring improvements to the existing methods and optimisation.

The work of this chapter and the data generated is published in Pattern Recognition Letters - Pattern Recognition and Crowd Analysis, 2013 [55], in IEEE Signal Processing Letters [8], in Computer Vision and Pattern Recognition workshop 2014 [54], and in the Sensor Signal Processing for Defence conference 2014 [7].

3.1 Introduction

A preliminary step to any computer vision application is that of feature extraction. The targeted visual features are defined by the availability and capability provided by the system, and the requirements of the application they feed into. Features range from very low level metrics such as optical flow [4] to High level object detection [30] [25], tracking [45] [42] [43], and speech recognition. Low level features such as optical flow are typically used in environments which hinder individual object tracking, such as dense crowds [39] [61] [76] [5]. See image 3.3.1 for illustration of optical flow. Our application uses a behaviour representation focused upon the actions and interactions of the individual human being, and we therefore require a method of feature extraction which provides information about the individual human. This requires some semantic a priori knowledge of our target appearance and motion to perform detection and tracking. A large amount of research has been carried out on the subjects of both detection and tracking. Both remain ongoing areas of intense interest with advances being regularly made. Given the amount of research carried out we utilise existing methods developed in recent years. We make some modifications upon the state of the art techniques to fit our purposes, particularly when we are not constrained by real-time processing we opt for a more robust but costly solution. We implement detection and tracking in C++ using mostly open source libraries, and use Matlab to implement head pose extraction.

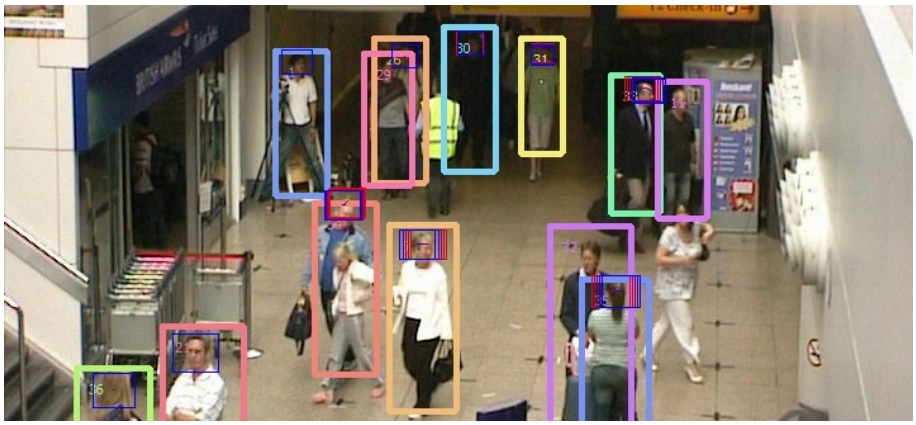


Figure 3.1: Illustration of Felzenszwalb DPM on the PETS 2007 scene 4 data. Detections are shown with a short bounding box track around the head to show local tracking.

3.2 Human Detection

At the time of our research the state of the art human detection technique was the Felzenszwalb DPM [30]. Although, subsequently, techniques based upon deep learning [33] have recently emerged which beat the performance of the DPM. We use the version of DPM from OpenCV 2.4.6 as the basis of our implementation. We modified the OpenCV DPM implementation by optimising the correlation calculation, multi threading the algorithm, and providing the capability to detect partial models for occlusion.

Deformable Part-based Models were first introduced by Fischler and Elschlager under the name pictorial structure models in 1973. The idea was brought successfully forward by Felzenszwalb in 2007 and when combined with modern machine learning took the form of Deformable Part-based Models in 2009. The DPM consists of parts and springs; the 'parts' are local appearance templates modelling the HOG appearance of the target. The 'springs' are a spatial prior modelling the connections between parts; allowing deformation of the model at a cost. The local appearance of an object is easier to model than the global appearance which does not allow for deformation. Furthermore the approach generalises to untrained configurations, requiring far less training for the range of possible global appearances. The DPM includes a more generic global appearance part amongst the numerous local parts at twice the image resolution of the global part. The detection is carried out on a pyramid of image resolutions in order to capture the range of pedestrian sizes in the image.

The DPM searches for detections, in this case people, within an image using the sliding window process. Each window is scored based upon the optimal configuration of parts $(p_1, ..., p_n) \in P^n$ given the spring constraints and the closeness of match of the query window to the parts. The optimal configuration is solved as a dynamic programming problem of complexity $O(nh^2)$. The score of a potential detection given a particular configuration in a window is given by:

$$S(p_1, ...p_n) = \sum_{i=1}^{n} m_i(p_i) - \sum_{(i,j)\in E} d_{ij}(p_i, p_j) \tag{3.1}$$

Where the score S is composed of the part matching scores m_i and the spring

costs for the configuration d_{ij}. The maximal response at each window location is calculated and a heat map generated of the window responses. Clusters of potential detections are then run through non-maximal suppression in order to return the singular positive matches that exceed a detection threshold.

3.2.1 Occlusion Handling

We provided an additional functionality which permitted the detector to assess only partial components of the HOG-based descriptor. When specified the detector reduces the parts it uses by excluding any parts which are outside of the reduction area. We achieve this by excluding a part of the DPM if optimal placement of the part exceeds the user specified region of interest on the main global appearance component. Furthermore the global HOG representation is clipped to only the user specified ROI preventing the model looking for a full representation of the target. The process reduces the detection scores returned from the classifier, accordingly the detection threshold must be matched in order to maintain the detections.

3.3 Human Tracking

The task of pedestrian tracking is the task of maintaining an accurate estimate of the location of a human through the entirety of that target's journey in the scene. Typically in surveillance the object motion is smooth, with no sharp changes, or jumps in the track. Fixed cameras are common, meaning that there are no sudden changes in the background. The appearance of tracked pedestrians changes gradually with a change in perspective and real world rotation. Due to quantisation of time, the continuous change in appearance is segmented into images, creating short discontinuities in which the sudden displacement of the tracked target and the change in appearance must be overcome. A sequence with an infinite frame rate would have no discontinuity and tracking would be trivial, however the lower the frame rate, the greater the discontinuous displacement in position and appearance, and the more robust the tracker needs to be.

We turn to the TLD tracker developed by Kalal [46] to solve the tracking task. The TLD algorithm consists of its three namesake elements: a tracking component; a learning component; and a detection component. These combine to create an algorithm that, when initialised on an object of interest, can track the object while it stays in frame and bootstrap learn the targets appearance in real-time. If tracking is lost, due to obscuration or exiting the scene then it can be re-identified by a sliding window detection process.

3.3.1 Implementation of TLD

We largely draw from the original work of Kalal [46] in this description of the TLD algorithm. The TLD algorithm is initialised with an image patch p which is re-sampled to 15x15 pixels irrespective of the original aspect ratio. The Object Model M built from the tracking process consists of positive p_m^+ and negative patches p_n^- such that $M = \{p_1^+, p_2^+, ..., p_m^+, p_1^-, p_2^-, ..., p_n^-\}$, where negative patches often correspond to background or hard classification cases. p_1^+ represents the first object bounding box which the algorithm was initialised upon. Given a image patch p there are several similarities to consider:

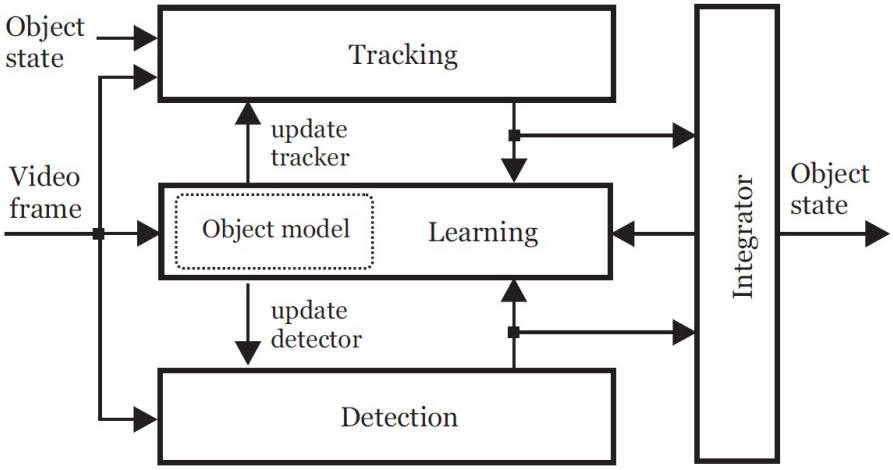


Figure 3.2: Block diagram of the TLD framework showing interconnection between the teacking, learning, and detection elements. Image originally from Kalal's work, [46]

- The Similarity with the positive nearest neighbour $S^+(p, M) = max_{p_i^+ \in M} S(p, p_i^+)$
- The Similarity with the negative nearest neighbour $S^-(p, M) = max_{p_i^- \in M} S(p, p_i^-)$
- Relative similarity, ranging from 0 to 1 signifying confidence in a match $S^r = \frac{S^+}{S^+ + S^-}$

S^r similarity resembles how much a patch resembles the object model and is used to define a Nearest Neighbour similarity. A patch p can be said to be positive if $S^r(p, M) > \Theta_{NN}$, otherwise negative. The classification margin is thus defined as $S^r(p, M) - \Theta_{NN}$ where Θ_{NN} is a configurable parameter giving a trade-off between accuracy and recall.

The object detector looks for potential cases of the target object with a scanning window on the image. All possible scales and shifts of the initial bounding box are assessed, with parameters:

- Scale step: 1.2
- Horizontal step: 10% of width
- Vertical Step: 10% Height
- Minimum bounding box: 20 pixels

Generating 50,000 bounding boxes for a 240x320 image. Thus, efficient classification of bounding boxes is required. The Nearest Neighbour classification is inefficient as it requires a search over all instances. Thus a cascade classifier is proposed consisting of three stages, each capable of vetoing a classification: patch variance, ensemble classifier, and nearest neighbour classifier. The initial light weight stage (patch variance) rejects windows which have a grey value variance less than 50% of the patch selected for tracking. Efficient calculation of variance $V_p = \mathbb{E}(p^2) - \mathbb{E}^2(p)$

with expected value $\mathbb{E}(p)$ calculated using integral images leads to a light weight preliminary step. This stage normally rejects 50% of negative patches. Non-rejected patches are passed to the Ensemble Classifier. The ensemble consists of n base classifiers. Each base classifier, indexed by i, performs a number of pixel comparisons on the patch resulting in a binary code $\mathbf{x}$, which indexes to an array of posteriors $P_i(\mathbf{y}|\mathbf{x})$, where $\mathbf{y} \in 0, 1$. The posteriors of individual base classifiers are averaged and the ensemble classifies the patch as the object if the average posterior is larger than 50%. We use an ensemble of classifiers defined by a d random pixel comparisons. First, the image is convolved with a Gaussian kernel to increase the robustness to shift and image noise. Secondly, each of the set of pixel decisions returns a binary result, which are concatenated into vector $\mathbf{x}$. Each base classifier has a distribution of posterior probabilities $P_i(\mathbf{y}|\mathbf{x})$ with 2^d entries. The posterior probabilities are estimated as $P_i(\mathbf{y}|\mathbf{x}) = \frac{\#p}{\#p+\#n}$ where $\#p$ is the number of positive image patches assigned the same vector $\mathbf{x}$ and $\#n$ the number of negative. Having filtered most of the bounding boxes in an image the remaining patches are passed to the Nearest Neighbour classifier. A patch is classified as the target object if $S^r(p, M) > \Theta_{NN}$, where $\Theta_{NN} = 0.6$. Kalal in his original work that sets the parameter empirically and finds that the value is not critical. He observes that similar performance is achieved in the range (0.5-0.7). Patches that are deemed positive at this point, after going through the NN classifier, are deemed true detections of the target object.

Object detection forms only one part of the Tracking Learning Detection algorithm. The tracking component of TLD is based on Lucas-Kanade Median-Flow tracker [62] extended with forward-backward failure detection. The objects translation between consecutive frames is estimated from the displacement of key points. The reliability of each key point displacement is calculated and the median reliable displacement is taken as the object displacement. Similarly to the original Kalal work we use a grid of 10 by 10 points.

In an alternative piece of work Kalal develops Forward-Backward tracking for the Lucas-Kanade tracker. We implement this technique here to catch tracking failure in the short term motion tracking. Tracking is performed forward and backward in time and the discrepancies between the two trajectories are measured. The error enables reliable detection of tracking failures and selection of reliable trajectories for an object. Let $S = (I_t, I_{t+1}, ..., I_{t+k})$ be an image sequence and x_t be a point location at time t. Using Lucas Kanade tracking algorithm [62] the point x_t is tracked forward for k time steps. The resulting trajectory is $T_f^k = (x_t, x_{t+1}, ..., x_{t+k})$ where index f signifies the trajectory is forward tracked.

The validation trajectory is first constructed. Point x_{t+k} is tracked backward up to the first frame and produces $T_b^k = (\hat{x}_k, \hat{x}_{k-1}, ..., \hat{x}_t)$. The Forward-Backward error is defined as the distance between these two trajectories:

$$FB(T_f^k|S) = \Delta(T_f^k, T_b^k) \tag{3.2}$$

The distance measure is arbitrary and should be designed to fit the tracking feature space available. We deviate from Kalal's work by using the Euclidean distance between all time corresponding points in T_f^k and T_b^k rather than the median distance between the start and end points:

$$\Delta(T_f^k, T_b^k) = \sum_{t \in k} \frac{||x_t - \hat{x}_t||}{k} \tag{3.3}$$

We designate a track as 'failed' when the FB error, equation 3.2, exceeds a threshold λ_{FB} which we set to 5 pixels. The exact value for λ_{FB} does not particularly

(a) (b)

Figure 3.3: Image (a) is an illustration of the TLD tracking upon head detections returned by the Felzenszwalb pedestrian detector on PETS scene 4. Tracking is the first stage of our algorithm pipeline. Image (b) illustrates the concept of optical flow. Green arrows indicate the pixel motion between two frames.

matter, but represents the strictness at which the tracker does not return a bounding box and relies upon the object detection to locate the track.

We initialise the TLD algorithm upon a detected pedestrians head bounding boxes from the previous detection phase. We track the heads of pedestrians rather than the full body because the head of a pedestrian is less commonly occluded from an elevated camera perspective. After initialisation the short term motion tracker makes the first motion estimates. We select the TLD tracker to extract pedestrian tracks in our work due to its robustness to occlusion, and particularly for its ability to discriminate between multiple similar targets using positive-negative learning. The original TLD tracker is tested in the work of Kalal [46], [45], [44], we refer to this work for more information. We next describe and test modifications we made to the TLD algorithm to suit our specific use.

3.3.2 Colour Space Re-Identification

Target re-identification is a persistent problem in the field of target tracking and identification [46] [40] [69] [67]. When a tracked target of interest is occluded by objects in the scene or leaves the scene entirely, it is often not possible to re-establish the track as belonging to the same target. Naturally, for target tracking systems to be of increased value, it would be advantageous to link together disjointed tracks into one continuous track, and therefore aid identification of that target and the users understanding of the scene. This will aid behaviour profiling which grows in strength as more information is gathered. Furthermore, due to a particular industrial application of our work re-identification of people is desirable. However more information cannot be given about this project. Currently, the TLD tracker used in our tracking work does not take into account the colour information of the target. Intuitively, the additional discriminative information characterising the targets clothing or physical attributes are unlikely to change through an occlusion. The key question we address here is whether or not colour information is of use when dealing with the problem of re-identification and if so, which colour space is best suited for

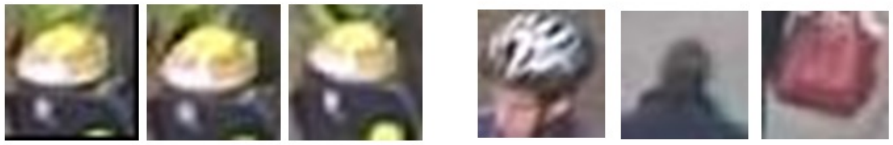
this task.

One popular tracking method commonly using colour information is mean-shift tracking [51] [21] [23] [22]. Mean shift tracking exploits the concept of the non-parametric density gradient estimator. It uses the colour histogram to model object probability density and moves the object region of interest in the largest gradient direction. Kumar [51] reports on the use of a tracker combined with colour information to improve object re-identification under occlusion. Motion tracking is achieved using mean-shift. A weighted colour distribution is maintained during tracking which helps discriminate motion by searching for the optimal translation to maintain the colour distribution. They present results from dense traffic data with 5-15 objects in the scene at any instant. Overall tracking accuracy is improved from 85.3% to 94.7% with the use of colour and motion. Chitaliya [21] presents a simple and fast block matching algorithm using a predictive motion vector for object tracking. The algorithm is enhanced using colour histograms for matching criteria for the motion tracking. Comaniciu and Meer [23] proposed a weighted colour histogram to represent the target object in an ellipse. The histogram is populated using the Epanechnilov kernel profile which weights the ellipse based upon centroid distance. Mean shift is used to find the location of target model in the current frame using the Bhattacharyya coefficient as a histogram distance metric. The benefit of the colour histogram is that has a high rotational invariance and is relatively unaffected by motion, and as such is a reliable metric for matching after occlusion. Additionally, kernel-based methods are computationally efficient, however, they do not encode positional information. Two objects may have similar colour histograms but have dramatically different appearances due to the distribution of the colours. For this reason we take an approach more similar to the colour correlogram [22] which encodes the positional information of colour.

Modifying the existing TLD algorithm allowed us to exploit the advantages of a well-documented and tested state-of-the-art tracking algorithm [42]. See 3.3.1 for a detailed explanation of the implementation of the algorithm. The TLD algorithm removes colour information by flattening an image to greyscale prior to processing. We modify the algorithm to accept three channel images instead of single channel images. The colour channels are unspecified, allowing different 3-channel colour-spaces to be tested within the algorithm. For detection in the colour space, the feature vector used to analyse each bounding box was expanded to include information from each of the three channels. Where originally grey pixels were compared as a binary decision in the ferns, single pixels from one of the three colour channels are compared and the feature vector extended, thus encoding colour information. Positional information is encoded as the feature vector stores pre-defined pixel co-ordinates for the image ROI which is always scaled to a uniform user defined size.

3.3.3 Experiment

Our hypothesis is that colour information is a profitable feature in the manifold representation of object appearance. We evaluate this hypothesis by investigating whether colour information can consistently provide an increase in true positive re-identification of a lost target. As the main focus of this investigation is re-identification improvements that could be gained from using colour information, it is sufficient to isolate the detection process of the TLD algorithm and manually train the model. In this way the impact of tracking, or tracking with colour does not obscure the impact of colour upon the TLD bootstrapped detector.



(a) (b) (c) (d) (e) (f)

Figure 3.4: (a) (b) (c) Samples of head images used in training in RGB colour space. (d) (e) (f) examples of negative images in training. All images have been enlarged to increase visibility.

For this experiment we use the Oxford dataset [12]. This dataset comprises a video of pedestrians walking up and down a high-street. We extract 5000 variable length tracks for each colour space (HSV, RGB, YUV and greyscale) to test upon. These colour-spaces were chosen as they are all in common usage and are all sufficiently different from each other to present different information to the algorithm. The same 5000 tracks were used across multiple colour spaces to enable a fair comparison, with the images converted to each colour space and indexed to ensure exact comparison between colour spaces. Each track consists of an equal number of positive and negative training images. The positive training images are a sequential set of head boxes taken from a single track in our data. The negative training images are of different head boxes and other head shaped objects in the data set.

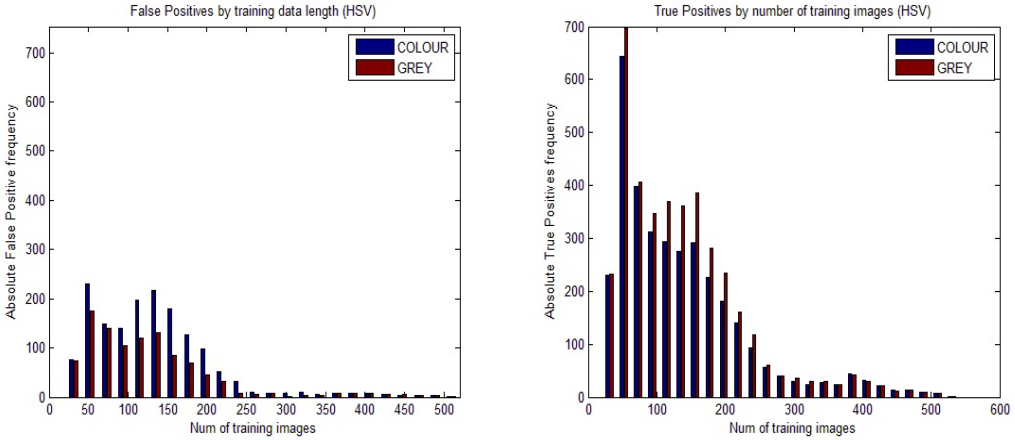
Having trained the algorithm on a specific track, the algorithm is then presented with 20 images, which it then scores on similarity to the track data it has been trained on. Out of the 20 test images presented to the classifier, one is true, 19 are false. The true track image is taken from a random number of frames in the future of the track ensuring there is a discontinuity in the track before the test image is selected. The gap ranges from 10 - 30 frames, with a uniform probability of selecting any number of frames in this range. The highest scoring image from the 20 possible images is taken as the classifiers selected image for re-identification.

It should be noted that the experiment is carried out on only the heads of pedestrians. One of the reasons for training and testing on a heads is that in a military context, it can be assumed that clothing is relatively uniform whereas heads will have variations.

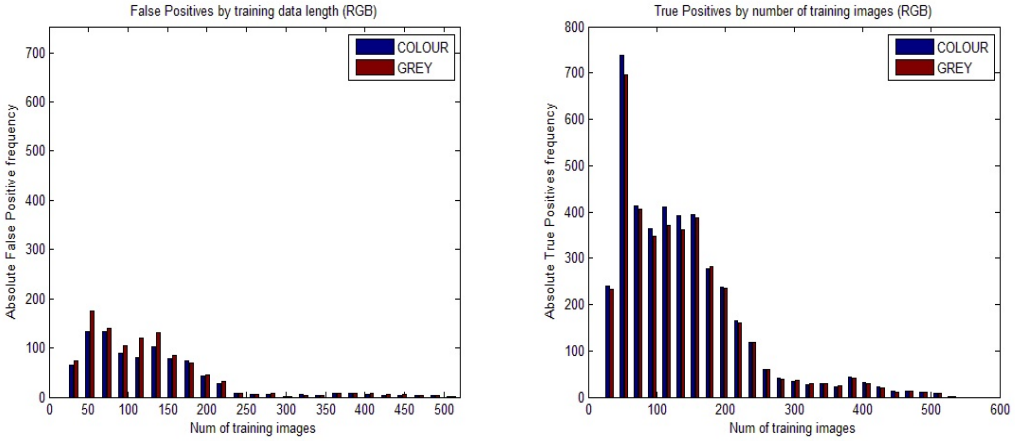
3.3.4 Results

The histograms, shown in 3.3.4 illustrate the true-positive and false-positive re-identification of the target binned by the number of training images. The absolute True Positive (TP) and False Positive (FP) frequency drops with the tests upon a higher number of training images. This is a result of using real world track data where shorter tracks are more common than long tracks. In our training and testing data there are only a few tracks that have a large number of training images above 250 frames. The colour and grey results can be directly compared only when they are in the same track length bin. This is because different tracks were used for each individual experiment, so variations between bins might be due to differing tracks, rather than solely the number of training images used. Thus any comparison between bins of the same histogram conflates multiple effects.

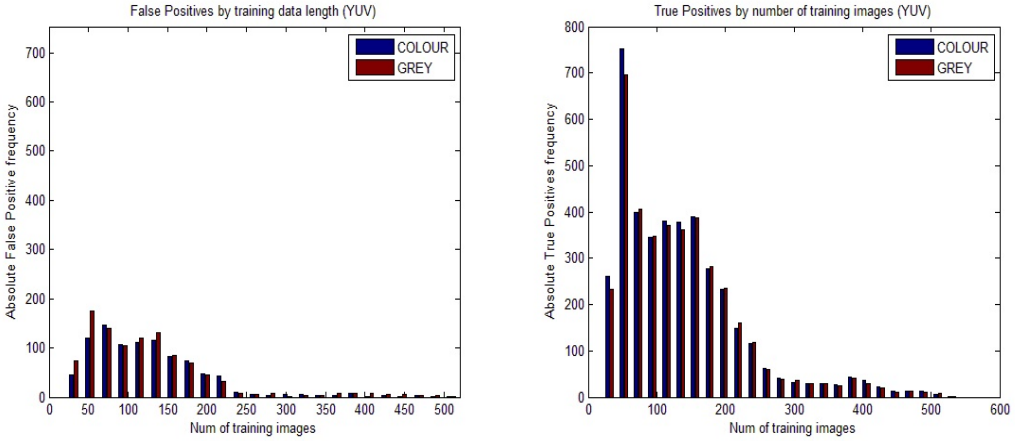
The results show that the use of colour information offers an improvement over the greyscale TLD algorithm in both the RGB and YUV colour spaces. However,



(a) HSV (b)



(c) RGB (d)



(e) YUV (f)

Figure 3.5: Six figures illustrating the impact of colour information upon true positive (TP) and false positive (FP) reidentification. HSV (a) and (b) show poor results when HSV colour is introduced. RGB shows a supression of false positives in (c). YUV shows similar suppression of false positives in (e) but to a lesser degree.

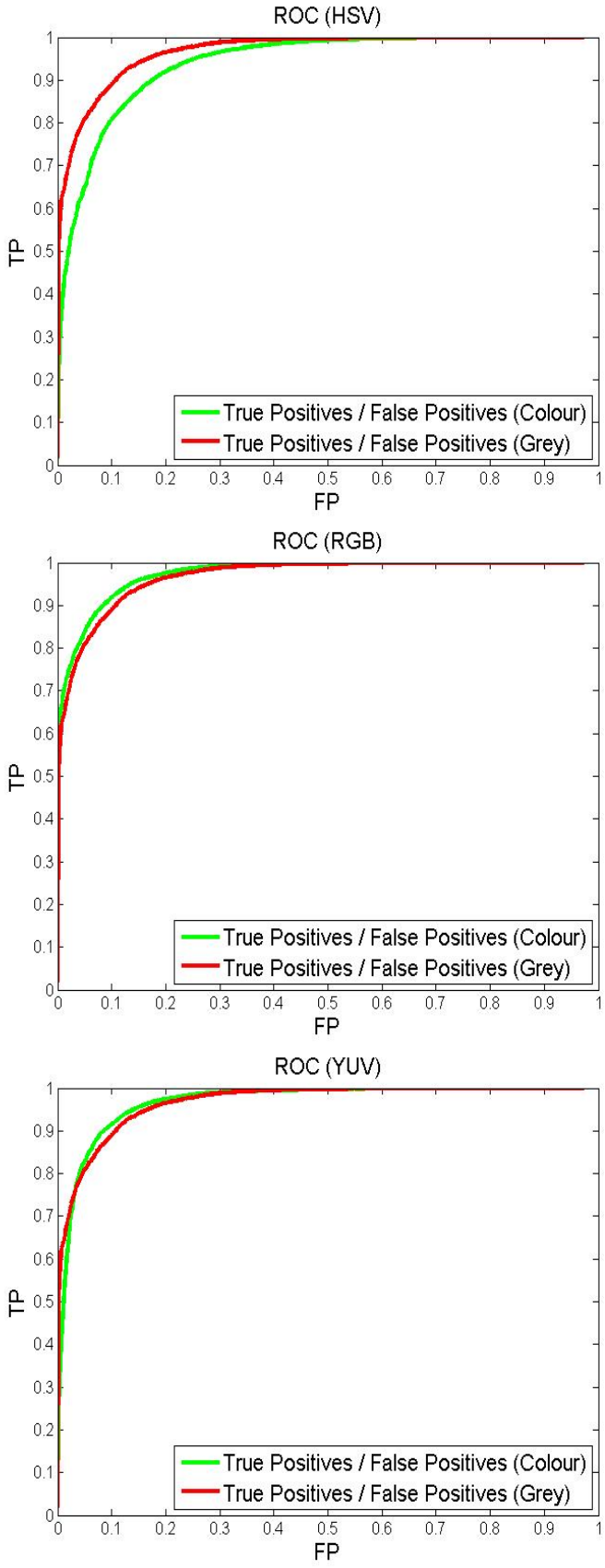


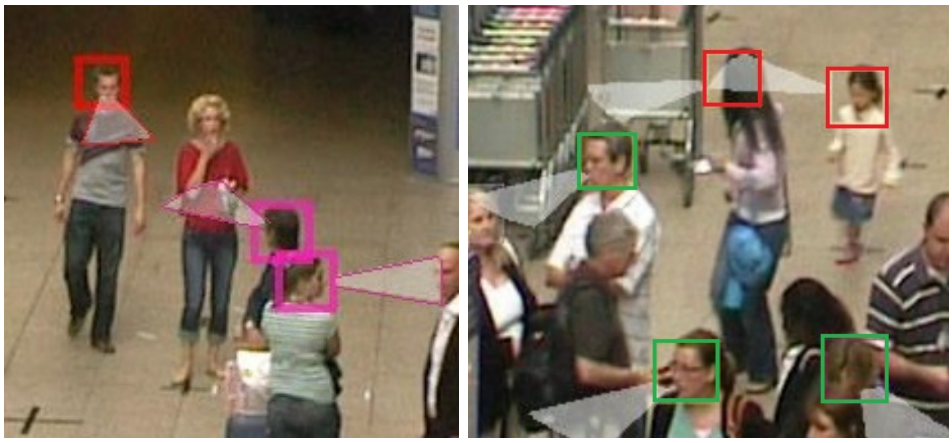
Figure 3.6: Illustration of the change in true positives and false positives for each of the colour spaces tested against. Ideally there is a separation between the true positive false positive rate for greyscale and colour, with the colour ROC (green) tending nearing 1, indicating an improved detection rate with the inclusion of colour. For both YUV and RGB introducing colour improves the re-identification classification. However it vastly decreases the confidence in true positives for HSV.

for the HSV colour space, performance is worse than the greyscale TLD algorithm in both TP and FP re-identification. For RGB, the confidence in true positives is increased by 4% and the confidence in false positives is reduced by 25% with the addition of colour information. Similarly for YUV colour space the confidence in TP re-identifications is increased by 1.3% and confidence in FP is reduced by 17.6%. The improvement in performance for RGB and YUV can be seen most dramatically in image (c) and (e) illustrating the reduction in false positives. Figures 3.3.4 (d) and (F) show an improvement in true positives. Figure 3.3.4 demonstrates the improvement in confidences resultant from the introduction of colour information. The implication is that it would make it easier to filter out false positives through the introduction of a threshold value, given that we saw a greater separation between the distributions of TP and FP results. For HSV the reverse is true, showing that the HSV colour space increases confusion between true and false positive examples. Given these results we proceed with tracking using a RGB enhanced version of the TLD algorithm in the tracking process for our surveillance data. The use of colour in the object model improves the robustness of our tracking allowing us to create longer tracks, and thus more complete behaviour profiles.

3.4 Head Pose

Our behaviour analysis method makes use of the visual interest people take in their surrounding environment. In order to determine visual interest we must extract the head pose of the target pedestrian. We use two methods to determine head pose; hand annotated ground truth and automated head pose estimation. We first hand annotate each head image at each frame in order to provide a baseline head pose direction from which the error of the automatic estimation can be calculated. Furthermore this provides the means to verify our behaviour analysis upon ideal data prior to testing on realistic data with error. Hand annotation was achieved by tracking a mouse pointer moved by the user to point to the current angle that the head was posed at, this was captured at 10 frames per second for a single person at a time. As the head angular velocity is highly constrained the head pose is particularly predictable in the short term, and hand tracking was found to be adequate, with occasional small latency of a couple frames when the head motion is more erratic. In total we groundtruthed 3 datasets; 2 from the PETS 2007 data, and additionally the Oxford Data. For the Oxford dataset approximately 70,000 head images where annotated. For the PETS scene 4 near 90,000 head images where annotated. The PETS scene 0 data entailed over 50,000 head images providing ample training and testing data.

To determine the visual interest a target has in a scene we must first estimate the target's head pose at each frame. Given the head pose we can then determine the likelihood distribution of visual interest given typical scene interest points and people within view. Our method for extracting head pose is identical to the work 'Unsupervised Learning of a Scene-Specific Coarse Gaze Estimator' [13] with the exception of the image classification factor. Benfold's work uses a randomized forest of ferns to learn typical relations between pixel triplets for a given head pose angle. The randomized trees were trained in a weakly supervised fashion with examples of each head pose class being fed through the ensemble of trees such that at each end node for each tree a distribution over every class is populated showing the probability that a head image reaching this node belongs to any given head pose class. We next outline the head pose extraction algorithm. Unless otherwise specified the algorithm



(a) (b)

Figure 3.7: An example of head pose classification using the Benfold method (a). The coloured bounding boxes [red, pink] illustrate a spatial grouping of the tracks. In (b) we illustrate a failure case where head pose has been incorrectly classified (red box) and correctly classified (green box).

for estimating head pose is the original work of Benfold [11].

For a sequence of head images $I = \{i_x\}$, where each head image is associated with a motion vector $\mathbf{v}^t$ containing ground plane speed and direction. We use a conditional random field model to combine several factors $C = \{C_T, C_F, C_\omega, C_I\}$ as a linear product:

$$p(\theta, \omega | i, \mathbf{v},) = \frac{1}{Z(x)} \prod_{C_p \in C} \prod_{\psi_c \in C_p} \psi_c(i_c, \mathbf{v}_c, \theta_c, \omega_c) \tag{3.4}$$

Where function $Z(x)$ is a normalising constant ensuring that $p(\theta, \omega | i, \mathbf{v})$ resides in the range 0 to 1 for image i and velocity of head in image of $\mathbf{v}$.The algorithm performs a costly inference step later requiring the possible head pose directions θ^t and angular velocities ω^t to be quantised into discrete states. We use a range of states from 4 to 32, each increasing the angular resolution and computational time for completion. Angular velocity ω^t is represented as a vector of three weights $w^t = (w^+, w^0, w^-)^T$ which define the expectation of the head orientation rotating positively, negatively and staying stationary. The values for w^t are defined empirically via expectation maximisation on training data. We know detail the four factors $i_c, v_c, \theta_c, \omega_c$ that compose the head pose estimation Conditional Random Field.

3.4.1 Angular Velocity

The angular velocity component weights for vector w^t for the distribution over ω^t are expected to be correlated with those for $w^t + 1$ because head angular velocity is physically limited and movements last several frames. We can represent the expected angular acceleration of the head by a 3x3 matrix A, predicting the angular motion w_{t+1} from w_t resulting in the factor function:

$$\psi_c(\omega_t, \omega_{t+1}) = (\boldsymbol{w}_t) A \boldsymbol{w}_{t+1} \tag{3.5}$$

Elements in A represent the probability of transitioning between different head angular momentums: positive rotation, negative rotation, and no rotation. The

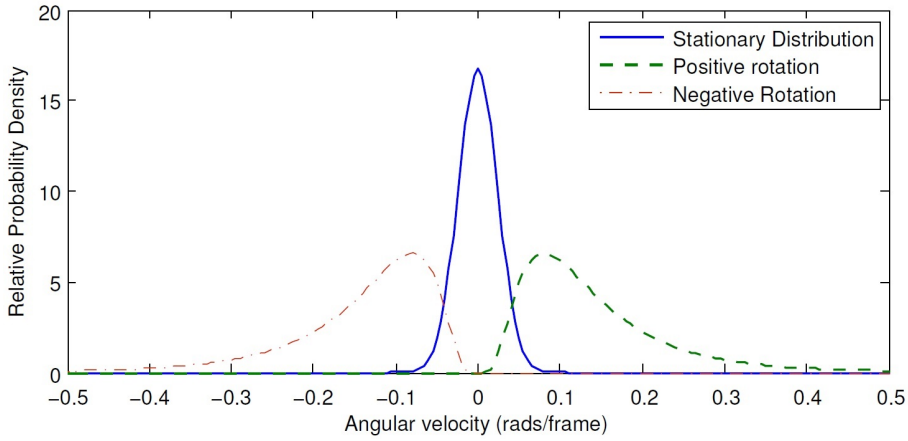


Figure 3.8: From Benfold's original work [10], this image illustrates the angular velocity components.

elements were estimated from model data by Benfold in his original work.

3.4.2 Image Classification Factor

This factor estimates head orientation from head images by classification into one of the quantised head pose states. We deviate from the Benfold work in our formulation of the image classification factor. Benfold uses a forest of randomised tree classifiers, with each leaf node containing a histogram over head pose bins. The tree classifier is fast and subsamples the data, a requirement for the video rate processing Benfold is restricted by. We implement an offline version of the algorithm, and for this reason we opt for a slower, yet exhaustive, representation of the feature space. Specifically we define a Matrix Y which holds the bi-pixel comparisons for all possible pairs of pixels in image i_x given every pixel in the image $P = p_1, p_2, ..., p_m$. The comparison between pixel p_n and pixel p_m for a particular colour channel returns:

$$\delta_{p_n,p_m} = \begin{cases} +1, & \text{if } p_n \geq p_m \\ -1, & \text{if } p_n < p_m. \end{cases} \tag{3.6}$$

This binary representation of a head image captures the global gradients, the gradient between each pixel and every other pixel, for each colour channel of a head image. Defining a head pose class simply takes the mean pixel comparison for any given pixel. Thus we can define the similarity $p(\theta|i_q)$ any query head image i_q, and binary matrix Y^q has to a head pose class θ as:

$$p(\theta|i_q) = \frac{1}{|N_\theta|} \sum_{n \in N_\theta} \prod_{p \in P} (Y_p^n Y_p^q) \tag{3.7}$$

We are thus defining the similarity of any head image to a head pose class as the mean Euclidean distance to all members of that class N_θ. The image classification factor is thus defined as $\psi_c(\theta^t, i^t) = p(\theta|i_q)$ where $p(\theta|i_q)$ is defined in 3.7. This form is computationally more expensive and exhaustively samples the set of all bi-pixel comparisons. We later evaluate the difference in classification accuracy for our method against Benfold's randomised forest of decision trees approach 3.4.5.

3.4.3 Changing Image Factors

The above factor, image classification, is based upon the assumption that similar head orientations result in similar images. This factor is designed to represent the correlation between head pose changing and the image change. Our image classification factor captures the mean appearance of a head pose class, thus representing the main contrasts in an image whilst smoothing out noise and individual variation. Although there is an expectation of substantial variance in appearance for members in a class, for a particular member looking in a particular direction over a short period of time, we expect appearance to be relatively uniform. We depart from the Benfold algorithm at this point, defining the distance between any two head images Y^n and Y^m as the ratio of dissimilar pixel comparisons to the total number of comparisons:

$$\phi(i^t, i^{t+1}) = \frac{1}{|P|} \sum_{p \in P} \delta(Y_p^n, Y_p^m) \tag{3.8}$$

Where ϕ ranges from 0 to 1 and $\delta(Y_p^n, Y_p^m)$ returns 1 if the pixel compared to for that colour channel is the same, else 0 if not. The probability that ω^t will be represented by the stationary component ω^0 is thus $\phi(i^t, i^{t+1})$, and the probability that it is represented by either of the rotational components ω^+ and ω^- is $1 - \phi(i^t, i^{t+1})$. In accordance with Benfold's work, we can thus define the changing image factor as:

$$\psi_c(\omega^t, i^t, i^{t+1}) = W^0 \phi(i^t, i^{t+1}) + (1 - W^0)\Big(\frac{1 - \phi(i^t, i^{t+1})}{2}\Big) \tag{3.9}$$

Where W^0 is the element of $\boldsymbol{w}^t$ corresponding to the probability of ω^t being represented by the stationary component.

3.4.4 Head Motion Factors

The final factor in the CRF is the head motion factor which encodes the probability of transitioning from one head orientation to another. Unlike Benfold's original work we empirically train a transition matrix upon training data. We train matrix T which represents our prior knowledge of how head poses should change between pairs of frames. This approach is a simplified method compared to Benfold's original method of parametrising T and fitting the transition probabilities to the model data using constrained optimisation. The fourth factor $\psi_c(\theta^t, \theta^{t+1})$ is thus simply defined by counting the number of transitions between each pair of head orientations in the training data and normalising by the total number of head images.

Having defined all four factors we can calculate the probability of head image i^t belonging to any 1 of the head orientation classes using the factors equation 3.4. Subsequently we can calculate the most probable sequence of head pose states given the factors $\psi_c(i_c, v_c, \theta_c, \omega_c)$ using a form of the forward backward algorithm. The resulting vector Φ_n for track n is the head pose sequence.

3.4.5 Evaluation

We use both groundtruth and automatic head pose estimations in our experiment. By taking the mean angular error (MAE) between the automatic estimated field of view and the groundtruth gazing direction we found that for the Oxford data we

achieved an automatic head pose estimation with MAE of 25.4 degrees compared to the groundtruth. For the more challenging PETS scene 4 data we achieved a MAE of 36.9 degrees. This represents a moderate estimated field of view offset from the true gazing direction. Our results are comparable to Benfold's results (MAE of 23.9) on easier the Oxford data [12]. Chen and Odobez achieve an angular error of 18.4 degrees [20] on the Oxford dataset. As of yet there are no published head pose statistics for the PETS 2007 dataset, however we consider our results to be particularly good given the far greater deviation from walking direction 5.2, and lower quality image data than the Oxford data.

We test the raw image classifier accuracy of the Benfold method and our method to validate our approach. The classifier is evaluated in isolation to the post-classification forward backward smoothing; this is to isolate the efficacy of the head pose feature representation alone. Comparison of methods is achieved by running the Benfold ferns classifier and our Euclidean distance classifier in section 3.4.2 in parallel and making a comparison to the ground truth for both classifiers. Rather than taking the class probabilities at each frame and feeding them into the forward backward smoothing, we instead take the highest probability class as the head pose estimate and compare this to the ground truth head pose. Note that as we take the maximum of the probability distribution over all classes as the predicted class, it is possible that one method may have a greater number of maximum probabilities coinciding with true head pose, and yet the other method has a distribution better fitted around the true class. The better fitting distribution would result in a better estimate after forward backward smoothing over all frames. Whilst this may be the case we still do not apply the FB smoothing as it is important to assess the performance of the classifier alone as other smoothing and prediction options, other than FB smoothing, may be desirable which take into account tracking trajectory and body pose [49, 20]. We reduce the comparison of both methods to the ground truth to a single Mean Absolute Angular Error θ^e calculated as:

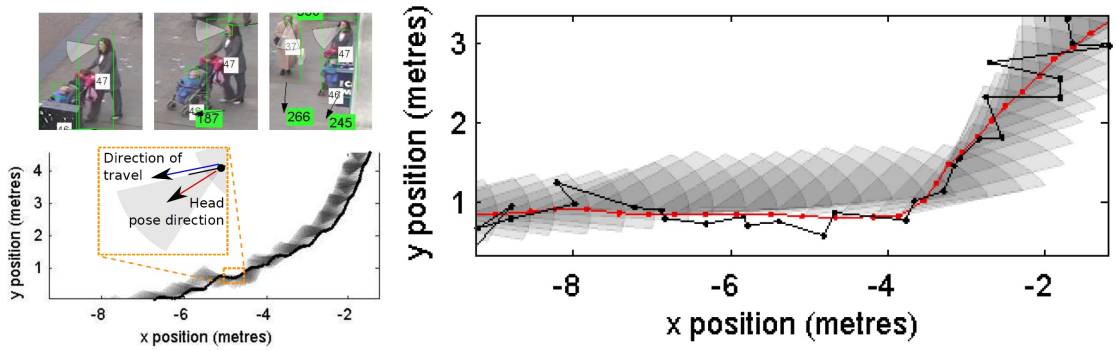
$$\theta^e = \frac{1}{T}\sum_{t=1}^{T} |\theta_t^G - \theta_t^E| \tag{3.10}$$

Where T is the total length of the head trajectory, θ_t^G is the ground truth head pose at frame t, and θ_t^E is the estimated head pose at frame t using either Benfold's or our method. The results obtained are as follows:

Table 3.1: Comparrison of raw head pose classification. We tested Benfold's fern-based method against our Euclidean method. For a small number of head pose classes (4 and 8) there was a negligible difference in accuracy, however we find for 16 head pose classes our methods has a reduction in error of 8.8% and for 32 head pose classes our method has a 13% decrease in error.

# of Head pose classes	Benfold: MAAE	Ours: MAAE
4	73.68	**71.30**
8	**66.11**	67.21
16	69.60	**63.47**
32	67.25	**58.51**
approximate time	30 mins	600 mins

We find that for a small number of head pose classes (4 and 8) there was a negligible difference in accuracy, however we find for 16 head pose classes our methods



(a) (b)

Figure 3.9: (a)(Top) Sample frames from the Benfold dataset showing the head detections & pose of a pedestrian. (a)(Bottom) The extracted ground truth trajectory and head pose behaviour of the person over time. (b) Part of a simulated track showing a 45 degree turn. The true target position is shown in red, observations in black, and head pose as grey sectors.

has a reduction in error of 8.8% and for 32 head pose classes our method has a 13% decrease in error. The findings indicate that our method has better fidelity however this comes at a large increase in computational cost. Our method took approximately 10 hours to train and classify 3 minutes of data, whereas Benfold's method took approximately 30 minutes when implemented in Matlab. Clearly our method is suitable for only offline methods, however, we have not explored optimization or implementation in a lighter language than Matlab.

3.5 Combining Head Pose with Motion Tracking

Recent advances in head-pose detection at a distance make it possible to incorporate head pose information as an intentional prior to human tracking. The Benfold model [13] allows heads to be detected and tracked and head-pose to be identified. Head pose information can be associated with trajectories of positions and we propose that head pose information should be used as part of person tracking algorithms. In the work of Baxter et al. [7] we develop and evaluate a novel approach to tracking using intentional priors. In this work intentional priors are introduced into a Kalman Filter (KF) to better predict pedestrian trajectories mediated by gazing patterns.

The Kalman filter provides an efficient recursive method of estimating the state of a system from a set of noisy measurements over time, where the seminal work can be found in [47]. As the basis for our tracker, we give a brief introduction to the relevant parts of the Kalman filter before introducing the components of our Intentional Tracker.

Kalman filter basics: Fundamentally, the Kalman filter attempts to estimate the state $\mathbf{x} \in \mathbf{R}^n$ of a discrete-time controlled process governed by the linear equation $\mathbf{x}_t^- = F_{t-1}\mathbf{x}_{t-1} + B\mathbf{u}_{t-1}$ with measurements $z_t = H\mathbf{x}_t + v_t$ (where t indicate time). w_t and v_t are the process and measurement noise and are assumed to be independent and normally distributed with zero mean and covariance Q_t and R_t, respectively. B is the process control input model and $\mathbf{u}_{t-1}$ is the control vector. We assume that B is the zero matrix so will not discuss it further and it will be omitted from later equations. Matrix F_t is often referred to as the motion or transition model

and relates the state of the process at $t-1$ to t. Matrix H is the observation matrix which we assume to be constant. The Kalman filter consists of prediction and update steps. The prediction step estimates the state of the system at time $t(\mathbf{x}_t^-)$ given all of the evidence prior to $t(\mathbf{x}_{t-1})$, and predicts the error covariance matrix P_t^-.

3.5.1 Integrating intentional priors

To integrate intentional priors into the Kalman filter we dynamically adjust the transition model F_t according to the intentional prior. Denote F_0 as the initial motion model. During the prediction step at time t we now generate a motion model I_t based on the intentional prior, and combine this with the initial motion model F_0 using a weighting component α. We will first present the generation of I_t for a head pose-based prior which assumes zero acceleration and has the general form:

$$I_t = \begin{bmatrix} 1 & 0 & d_t \cos(\theta_p) & 0 \\ 0 & 1 & 0 & \sin(\theta_p) \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.11}$$

Where d_t is the geometric distance travelled by the target between $t-1$ and t and θ_p is the predicted direction of travel based on the velocity $\mathbf{v}_t = [v_x, v_y]$ and estimated head pose deviation $\theta_d : \theta_p = \arctan 2(v_y, v_x) + \theta_d$, where θ_d is is assumed to be normally distributed: θ_d $N(\mu, \sigma)$ with parameters learnt from the scene and $\arctan 2$ is the 4-quadrant arctangent function. Having derived I_t we use weighting component α_t to combine I_t and F_0 as follows:

$$F_t = (1 - \alpha_t)F_0 + \alpha_t I_t \tag{3.12}$$

Intuitively α should increase in line with the strength of the intentional prior $\hat{s}_t$, where $\hat{s}_t$ combines magnitude and persistence. This can be achieved using a sigmoid function with optimal parameter values γ and τ derived via an optimisation procedure. The γ parameter adjusts the gradient at which the function moves from zero to one, while τ shifts the sigmoid along the x-axis. The resulting function can thus be adjusted to change the weight given for zero strength as well as the gradient at which the weight changes.

$$\alpha = \frac{1}{1 + \exp(-\gamma(\hat{s}_t - \tau))} \tag{3.13}$$

To calculate $\hat{s}_t$ we use the absolute magnitude of the deviations for the last 10 time steps. To eliminate small fluctuations in deviation/detection inaccuracies. We use a binning procedure to partition the velocity and head pose into 8 bins (numerically numbered 1:8), where each bin represents a 45° sector. The signal strength at time t is thus calculated as follows:

$$\hat{s}_t = |\sum_{k=t-10}^{t} Bin(\theta_k^g) - Bin(\theta_k^v)| \tag{3.14}$$

Where θ_k^g is the head pose direction and θ_k^v is the direction of travel. Having finally defined all of the components required to generate F_t, the remainder of the Kalman filtering algorithm remains the same.

Table 3.2: Percentage improvement of log likelihood using the 'Intentional Tracker' on real data from the Benfold dataset. Turn exemplars 1:3 have approximate trajectory changes of -90%, -40%, and -45% respectively.

Trajectory	Ex. 1	Ex. 2	Ex. 3	Mean
Turn	18.50%	9.18%	16.33%	14.67%
No - Turn	16.51%	12.59	15.28%	14.79%

We have showed that head pose and direction of travel are well correlated and provided statistical evidence that the intuition people look where they are going is true. The results of our pedestrian tracking experiments confirm that the intentional tracker is able to outperform the Kalman filter by as much as 23.61% on the simulated sample trajectories by means of reduced MSE. We also demonstrated performance on a sample of real pedestrian trajectories from the Benfold dataset [12], where the tracker achieved a mean improvement of 14.73% in log likelihood.

3.6 Conclusion

This chapter brought together research carried out in recent years in feature extraction from video. We detailed and illustrated the techniques we use to provide the data to our behaviour analysis system. We combined methods taken from object detection, target tracking and head pose estimation to build a set of algorithms capable of automatically tracking pedestrians and estimate their head pose. Where necessary we have enhanced the capability of the tracking, testing our hypothesis, and validating our enhancements to improve tracking accuracy. These improvements reduce tracking noise and thus feed forward to increased accuracy in our anomaly detection system. In summary, the work presented in this chapter made the following contributions:

- Integrating the intentional prior of head pose into pedestrian motion tracking
- Validating that colour information improves the TLD tracking algorithm, and determining which colour space provides the greatest improvement
- we propose and validate an alternative head pose classifier within the Benfold head pose estimation framework which has higher accuracy at increased computational cost
- Minor additions and optimisation of the Deformable Part Based Detector, OpenCV implementation. We increase its capability to detect people under partial occlusion

We next present a motion-based human behaviour anomaly detection system using the motion data provided in this chapter.

Chapter 4

Context Aware Motion Behaviour Analysis

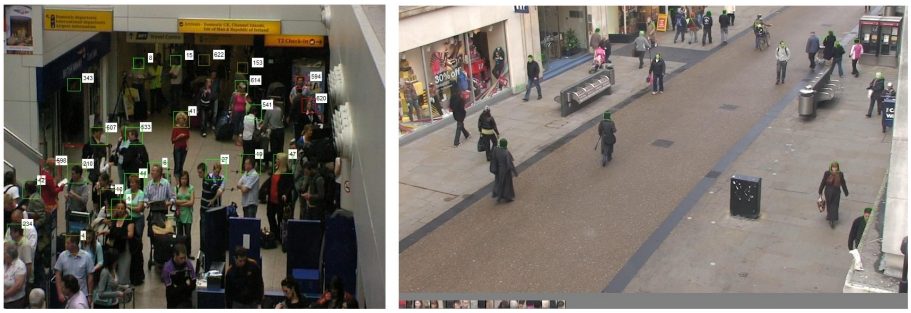
In this chapter we draw upon the work in the previous Chapter 3 to provide a system which leverages contextual information to improve the interpretation of behaviour and ultimately better find human behavioural anomalies in surveillance. We model human behaviour as a two part distribution containing a motion element which characterises the shape of the behaviour, and a context element which provides additional information separating subtle anomalies from the normal motion of behaviours. We use the data extracted in the previous chapter to demonstrate our method in 4 different surveillance scenes. We show that using an estimation of social connections in a scene, social context, and region classifications, scene context, we can improve behaviour anomaly detection. We evaluate our approach on real surveillance data and discuss the impact automatically generated contextual information has upon automatic surveillance. This chapter addresses our research objectives 4.1, 4.2, and 5.1; see section 2.7.

The work of this chapter and the data generated is published in Pattern Recognition Letters - Pattern Recognition and Crowd Analysis, 2013 [55].

4.1 Introduction

We previously established the intuitive notion that contextual information provides additive information upon which behaviour analysis can be enhanced. With this work we demonstrate the significance of two forms of contextual information; inferring social links between people in a surveillance and segmenting a heterogeneous surveillance scene. In doing so we provide a novel social strength metric, and provide validation of the growing trend in automatic scene understanding. Furthermore we demonstrate a novel social context-based anomaly detection procedure. This work further motivates the approach of using contextual information, and establishes two sources of information in human surveillance as a means to better interpret human behaviour.

We define an abnormal event in surveillance as one which has a low statistical representation in the training data [60]. Our approach is motivated by this definition with an emphasis upon contextual information as a method of creating separation between otherwise only subtly distinct behaviours. A good behaviour representation should encode the dataset in such a way that homogeneous clusters of behaviour can be segmented from the heterogeneous mass of data. Equally a poor behaviour representation is incapable of measuring the distinction between desired subgroups



(a) (b)

Figure 4.1: We illustrate here the tracked dataset PETS-2007 (a), and tracked Oxford data (b). The PETS-2007 data presents a challenging crowded environment and contains far less structure in the apparent motion of individuals in the scene. In contrast the Oxford data contains very structured trajectory information, and is sparsely populated. Our social context extraction is geared towards crowded scenes such as the PETS-2007 data, however this presents a harder surveillance challenge.

of data. Subtle behaviours provide a greater challenge because the information required to segment them from the greater set is not salient. Subtle behaviours can be handled in the following two ways; firstly by measuring more relevant information which better segments the data into homogeneous subsets, or secondly by implementing a better suited model which is capable of fitting the nuances of the data domain. In this research we tackle the former point; inspired by work in Scene Modelling [63] and Social Signal Processing [24] we demonstrate the extraction and use of high level surveillance information which provides a *contextual* basis to identify subtly abnormal behaviour. Simple surveillance scenes may not contain much contextual information, in fact at its simplest a surveillance scene can be said to have only one contextual state. In such cases a simple trajectory matching algorithm may be appropriate to detect outlier behaviour. However, a dynamic or crowded surveillance scene may be heterogeneous, and thus behaviour in one context may not be representative of behaviour in a different context. In any non-trivial surveillance scene contextual information such as scene region, social context, periodic events, and entry or exit points impact the dynamics of behaviour [52]. We can use this contextual information to provide further means of segmenting abnormal behaviours from the mass of data, and perhaps provide the means to segment subtle behaviours from the mass of data. For a more general discussion on contextual anomaly detection see [17] [77].

We evaluate our systems capability to detect *subtle* behavioural anomalies within a complex and crowded human surveillance scene. Our main contributions in this work are a novel method of acquiring scene structure information in surveillance, the development of a novel mutual information social group metric, and the demonstration that social and scene contextual information is effective in combination at anomaly detection.

Much of the literature relevant to this work has been reviewed in Chapter 2. The rest of this chapter gives a detailed account of the approach used to validate our hypothesis and sets out how the experiment will be carried out. Subsequently, the results are collected and discussed at the end of the chapter.

(a) (b) (c)

Figure 4.2: An example of social grouping from the Oxford data (a) and the PETS-2007 data Scene 04 (b) derived using our social connection strength metric in Chapter 4.4. Both (a) and (b) show a true positive result. (c) demonstrates a failure mode.

4.2 Feature Extraction

The extraction of pedestrian trajectories from surveillance video is non-trivial, particularly when there is occlusion and crowding. It is not our goal to develop a novel low level feature extractor and for that reason we rely upon the large amount of research in computer vision already devoted to producing tracking solutions. The methods we use to extract human motion in video is covered in detail in Chapter 3. We reiterate here to clarify our methodology. The extraction of pedestrian trajectories requires two main stages: detection of pedestrians, and tracking of targeted pedestrians. Detection is achieved using the Felzenszwalb part-based detector [30]. Tracking of human targets in the image plane is achieved with the use of the Predator TLD tracker [42]. We track the heads of pedestrians in the crowded PETS-2007 scene, see Figure 4.1 (a). for the second dataset, the Oxford data, we use the published tracking results provided by Benfold [12]. We select the TLD tracker due to high performance amongst state of the art trackers [43] and utilise its capability to learn a target model and discriminate between potential targets in a crowded surveillance scene. The pedestrian tracking performance of the TLD tracker is extensively tested against alternative recent tracking procedures in the author's paper [43]. Furthermore we enhance the TLD tracker with colour information which we find better suppresses false positive detections, see Chapter 3.

4.3 Scene Context in Surveillance

Building upon the work of Makris [63] our scene model consists of four potential regions: Traffic lanes, idle areas, convergence/divergence regions, and general area. Convergence and divergence is synonymous as there is no temporal direction. Each region is defined to isolate a different dynamic of a scene, and is captured as a relation between the direction, speed, persistence (the number of frames a trajectory last for), and energy and entropies of trajectories through the scene. For each of the four potential regions a heat map is constructed on the ground plane and a threshold segments positive regions from negative. Scene regions are mutually exclusive of each other. We define each of the four scene context regions as follows:

Traffic Lanes: A traffic lane represents an area of the scene which contains a high number of trajectories in a structured motion. The traffic region is defined as:

$$T_{xy} = \frac{N_{xy}}{\bar{N}} \frac{1}{-\sum P(\theta_{xy}) log(P(\theta_{xy}) + \frac{1}{\pi} \sum \sqrt{(\theta_{xy} - \bar{\theta}_{xy})^2}} \tag{4.1}$$

Where θ is a histogram of directions populated by all target trajectories to go through region x, y in the scene. The numerator N_{xy} gives the number of trajectories through the location x, y, and $\bar{N}$ gives the mean number of trajectories for any given location. High scoring traffic locations coincide with regions displaying a high number of trajectories, low directional entropy and low trajectory energy.

Idle Regions: The idle region captures the area of the scene which hold enough evidence of near stationary trajectories that the region is considered a legitimate place to remain idle.

$$I_{xy} = \frac{T_{xy}}{\bar{T}} \frac{v_{xy}}{\sum \sqrt{(v_{xy} - \bar{v}_{xy})^2} + \sum v_{xy}} \tag{4.2}$$

The mean temporal persistence T_{xy} provides the mean numbers of frames that trajectories persist for in the region x, y, this coefficient is balanced by the denominator $\bar{T}$ the mean number of frames for all regions. The speeds of trajectories observed in location x, y is denoted by histogram v. We define likely idle regions as those with a high mean temporal persistence, low speed and low speed energy.

Convergence Divergence areas: These areas of the scene are responsible for imposing a force which brings trajectories together or releases them allowing them to diverge. Typically such regions are appended to the ends of a traffic lane.

$$C_{xy} = \frac{\frac{1}{\pi} \sum \sqrt{(\theta_{xy} - \bar{\theta}_{xy})^2}}{-\sum P(\theta_{xy}) log(P(\theta_{xy})} \tag{4.3}$$

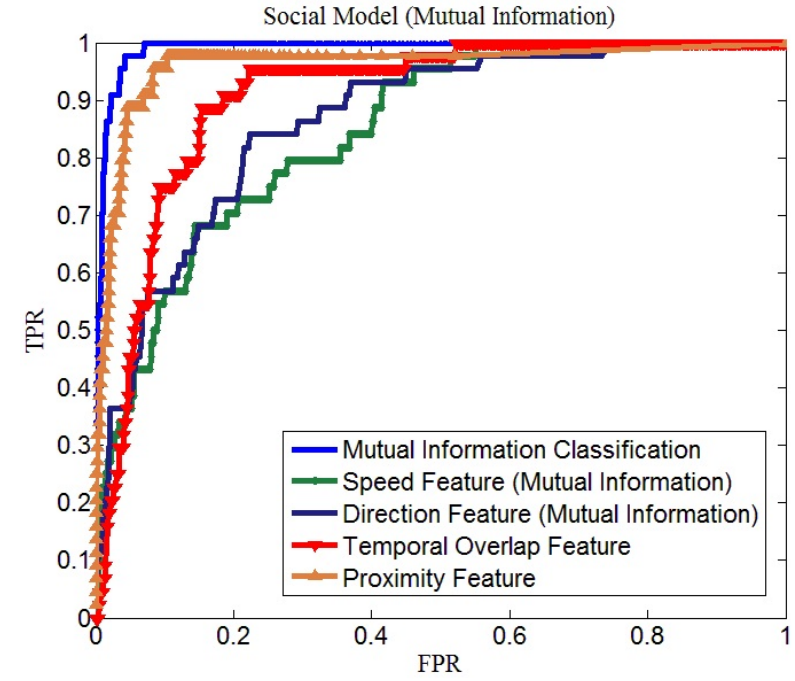
Where θ is the histogram of direction observed at x, y. We define the convergence region by a high directional energy low directional entropy region. Thus a structured splitting of trajectories over a region would be considered a likely candidate for a convergence or divergence region.

General Area: having scored the scene with the above region definitions we normalise the region intensity maps between [0,1], and apply a threshold to segment active regions. The remaining area of the scene not classified as any of the above regions is considered the general area. The interpretation of the general area is as the region which does not impose any influence on the motion vector of tracked pedestrians.

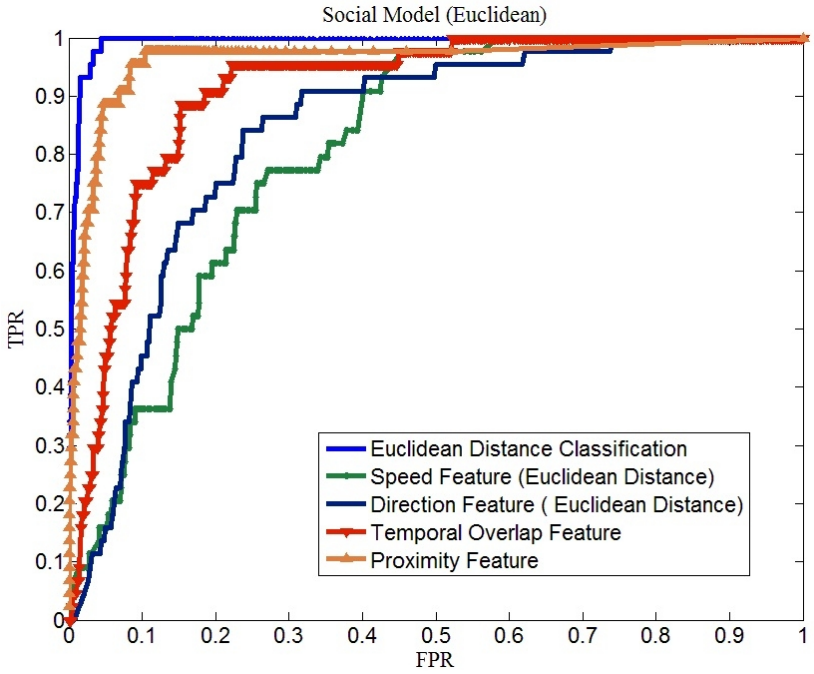
4.4 Detecting Social Dependency

The basis of our social model is the premise that a high degree of shared trajectory information implies a social dependence between two individuals. Our social model is geared towards effective detection of social groups in a moving crowd. Crowded surveillance provides an environment in which socially connected individuals are more likely to move together, and thus display more similar trajectory information. The more entropic the underlying motion of the crowd is the more salient similar trajectories will be. For an illustration of typical social pairs see Figure 4.2 (b).

We use a novel metric to identify the strength of pair-wise social connections consisting of the weighted product of multiple features. We identified 4 features as effective at detecting pair connections between two individuals: the mutual information of direction ($I\Theta_{ijt}$), the mutual information of speed (IV_{ijt}), the proximity



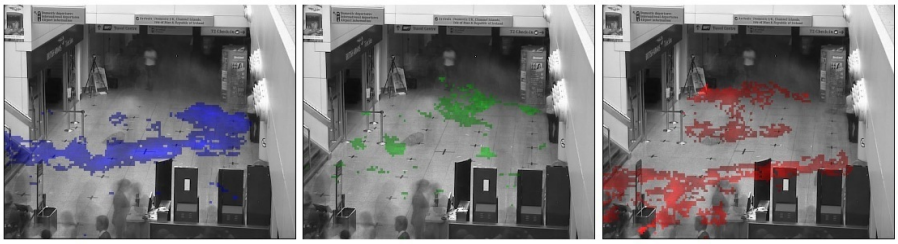
(a)



(b)

Figure 4.3: A comparison of the features which comprise the Mutual information social model (a) and for comparison the Euclidean distance equivalent (b) both trained upon the PETS 2006 dataset and tested upon the PETS 2007 data set. The proximity and temporal overlap in both metrics are identical. The critical difference is in the speed and direction information. We observe that the mutual information speed and direction metrics outperform the Euclidean distance feature metrics individually, however in overall true positive classification the Euclidean approach reaches a more optimal result

between two individuals (ΔP_{ijt}) and the temporal overlap ratio between two individ-



(a) (b) (c)

Figure 4.4: (a) (b) and (c) illustrate the automatic scene segmentation we arrived at using the all trajectories from the PETS-2007 datasets. Each unique scene context is designated by a colour; Idle region - Red, Traffic region - Blue, and Divergence region - Green. Areas of the scene not included in either scene region class do not have sufficient supporting evidence to be classified and as such remain blank.

uals (τ_{ijt}). We train a set of weighting variables $\alpha_{\Delta P}, \alpha_{IV}, \alpha_{I\Theta}, \alpha_{\tau}$ which weight each feature in the social metric based upon the classification score of each feature independently on the ground truth training data. The feature weights are distributed proportional to each feature's classification score. The features which compose the pairing metric are defined as:

$$\Delta P_{ijt} = \alpha_{\Delta P} e^{-\frac{\frac{1}{N}\sum_n |S_{it}-S_{nt}| + \frac{1}{N}\sum_n |S_{jt}-S_{nt}|}{2\ |S_{it}-S_{jt}|}} \tag{4.4}$$

For two tracked individuals i and j at frame t where S_{ij} is the distance between trajectory i and j at time t. The proximity between any two individuals ΔP is scaled by the distance between i and j to the set of all other individuals N in the scene. Thus we incorporate a measure of scene density which places a bias upon pairs being closer together in denser areas, and allows pairs to drift apart in sparse areas.

$$\Delta \tau_{ijt} = \alpha_T e^{-\frac{|T_i - T_j|}{2T_{ij}}} \tag{4.5}$$

Where τ_{ijt} is the temporal overlap ratio between i and j up to the current frame t, which is to say the ratio of time both individuals have existed contemporaneously to total time of existence, thus rewarding individuals who enter and exit the scene at similar times. T_i, and T_j is the frame length of trajectory i and j respectively, and T_{ij} is the number of frames in which both i and j have coexisted.

Whilst ΔP_{ijt} and $\Delta \tau_{ijt}$ are direct measures of trajectory statistics it is important to note that both $IV_{ijt}, I\Theta_{ijt}$ are more complex in nature. We use mutual information (MI) instead of the Euclidean distance as it handles non-linear and non-Gaussian random variables effectively and provides a principled method of comparing orthogonal feature dimensions. We define the Gaussian distributions of speed $P(v)$ and direction $P(\theta)$ as the Maximum Likelihood Estimation (MLE) derived from the most recent 1 second of trajectory data. The joint probability is calculated as the MLE Gaussian for the combined data of both person i and j over the last second. The mutual information between individual i and j is calculated for a number of temporal offsets thus permitting an individual reaction time to the trajectory it has dependence upon. Thus we calculate the mutual information between each individual with set time offsets of 10 frames consecutively forwards and backwards, and

take the maximal mutual information for all time offsets.

$$\begin{aligned} IV_{ijt} = &- \alpha_{IV} \sum_{b} P(v^i(b)) log_2(P(v^i(b))) \\ &- \alpha_{IV} \sum_{b} P(v^j(b)) log_2(P(v^j(b))) \\ &+ \alpha_{IV} \sum_{b} P(v^{ij}(b)) log_2(P(v^{ij}(b))) \end{aligned} \tag{4.6}$$

Where v^i is the MLE distribution over speed for person i over the most recent time window. The mutual information calculation for direction $I\Theta_{ijt}$ is structured identically to the above, replacing the MLE speed distribution v^i with the MLE direction distribution θ^i.

Each feature is used independently to classify pair connections between tracked individuals and scored with against the ground truth classification. We observed that the features of proximity between two individuals (ΔP) and the temporal overlap ratio between two individuals (T_{ijt}) present a significant ability to classify pairs in the test data. The overall performance is improved with the inclusion of the mutual information measures for direction and speed, see Figure 4.3. Whilst the individual features of mutual information speed and direction provide better classification we find there is a lack of correlation with the true positives exemplified by the Euclidean features of proximity and temporal overlap in this dataset. In this dataset the impact is a slightly reduced true positive rate. However we select the mutual information metric over Euclidean distance as it is a more principled method and scores better than the Euclidean features.

To measure the overall social connection strength between two individuals we utilise the pairwise strength in the previous step in the following way. A trajectory of length T frames consists of T tuples (S, v, θ) for 2D ground plane position S, speed scalar v and direction of trajectory in radians θ. We can calculate the pair strength at frame T between any two individuals i and j, for $i, j \in N$ where N is the set of all individuals in the scene for all frames. The social connection strength κ between two individuals i and j at time T is:

$$\kappa_{ijt} = \frac{1}{T} \sum_{t}^{T} IV_{ijt} I\Theta_{ijt} \Delta P_{ijt} \tau_{ijt} \tag{4.7}$$

$\tau_{ijt}, IV_{ijt}, I\Theta_{ijt}, \Delta P_{ijt}$ are the temporal overlap, mutual information for speed, mutual information for direction and proximity difference between person i and j, as detailed in the feature equations (4.4), (4.5), (4.6). We classify the social state S, for $S = \{0, 1\}$, by applying social strength threshold λ which is set empirically from the training data. Connections between individuals which score higher than λ are considered socially connected, providing the binary social context state used in the anomaly detection stage.

4.5 Detecting Anomalies

Anomaly detection splits into three distinct segments: the *behaviour representation*, the method for *calculating normality* of observations, and the algorithm for *detecting anomalies*.

4.5.1 Behaviour Representation

We represent the instantaneous state of behaviour with a four part feature vector $\mathbf{x} = \Re^4$, consisting of a bivariate motion component [speed, persistence], and the two contextual states [social state, scene region]. Speed is measured in meters per second on the ground plane, and social state is a binary state describing whether the individual is part of a social group or not. The persistence of an individual is a measure in frames of how long an individual has remained in the scene. Lastly, the scene region identifies the scene context region in which the individual resides, denoted by a numerical identifier. For an individual with trajectory length T frames we have T feature vector observations. The observations are accumulated to a discrete 4 dimensional feature space representing a 4D histogram, termed the behaviour profile X_i, for individual i. Defined in this way X_i consists of a feature distribution from a large number of observations. The advantage to this is that it hides short-term measurement noise resulting in a behaviour representation which is more robust. Furthermore, as measurement noise is often correlated rather than Gaussian white noise, the order independent nature of the behaviour profile X_i overcomes the appearance of anomalies that arise from structured noise. Our behaviour profile provides flexible temporal scaling of behaviours; something Dynamic Bayesian Networks struggle with, however it results in the loss of time series information which may reduce the descriptive capacity of the representation.

4.5.2 Normality of behaviour observations

As our approach is unsupervised, anomalies are discovered due to their contrasting nature to previously observed behaviour. Much work to date has focused upon a frequency-based analysis to determine the normality of behaviour observations. However, frequency-based anomaly detection suffers under the following assumption: that the normality of any observed behaviour is proportional to the relative frequency of observations of the behaviour. Whilst we can expect abnormal events to be rare, it is not the case that normal events are all frequent, and proportionally represented. We wish to distinguish here between the *normality* of a behaviour and the *expectation* of a behaviour. The expectation of a behaviour is how likely it is to occur next, whereas the normality of a behaviour is how permitted the behaviour is in the scene; how legitimate it is. A frequency-based analysis reveals expectation of each behaviour to occur next, not the intrinsic normality of the behaviour itself, thus missing the mark. We instead implement a Nearest Neighbour method to search for supporting evidence for an observation from others within the data. The normality of any behaviour is based upon its distance to the nearest K instances of supporting evidence *n*ot the frequency of observation for that behaviour. It does not matter how many people represent the behaviour, rather how well the K nearest behaviours represent the person in question. The advantage to departing from the frequentist approach is that we measure the degree to which something is an outlier rather than measure the local density. This is computational more expensive as we typically have to search over a larger area, however, in a sparse environment, it provides a method of measuring the distance of a observation when there are no other local supporting examples. There would otherwise be no way of determining whether a data point lies just outside of a cluster or very far out of a cluster. As we are most concerned with identifying the outliers, rather than the clusters, the non-frequentist K-nearest neighbour approach meets our needs.

Whilst a nearest neighbour approach could be expected to segment out anomalies

(a) (b) (c)

Figure 4.5: Illustrated here is three examples of anomalies detected by our system in the PETS 2007 data set. (a) shows two true positives with a false positive in the bottom left corner. The anomalies in (a) refer to anomaly Id: 6 and 7 in Table 1. In (b) two examples of loitering are detected, anomaly Id: 11 and 12. In (c) loitering is detected, Anomaly Id: 9, and 10.

with strong contrary motions, a subtle anomaly may not be distant from the set of normal behaviour with regard to the majority of features. A subtle anomaly may be abnormal for only a subset of features, and furthermore only when seen in the context of another feature. For example the speed is abnormal only when seen in the context of a specific scene region, rather than the speed and scene region both being independently abnormal. As such we need to assign a normality score to each feature in context of each other feature, independently of every other feature, a step critical to detecting subtle differences between behaviours. This step enables us to see context dependent distinctions between behaviours which when viewed in the full feature space are too subtle to impact a distance calculation. To represent each feature in the context of another we reduce our 4D histogram feature space to a set of 1D feature distributions $Y_n^{f1,f2}$ detailing the distribution of feature f1 given the currently observed value for feature f2 for person n at frame t. For a feature vector $\mathbf{x}_i$ with dimensionality D there are $D^2 - D$ feature context pairs covering each $\{f_1, f_2\}$ feature pairing, when $f_1 \neq f_2$. In our 4D feature space 12 individual feature pairs are assessed at each frame for each individual, each representing a different observation given context pairing. To reduce the dimensionality of X_i to 1 for a particular feature context pair we sum the distribution X_i for all dimensions f in the set of dimensions F where $f_1 \neq f_2$ resulting in a 2D joint distribution Y_n of observation feature f_1 and context feature f_2. We then take a further step reducing the 2D distribution to the target 1D distribution by taking the distribution through the current context feature value $f_2(i)$ only. Thus our resulting distribution $Y_n^{f1,f2}$ details the distribution of observed feature values for observation feature dimension f_1 given the context feature state $f_2(i)$. An example of which would be the distribution of the speed feature given the scene feature of idle region.

We apply the Nearest Neighbour (NN) function to distribution $Y_n^{f1,f2}$ and the set of all distributions Y to determine the nearest neighbour $Y_m^{f1,f2}$ to $Y_n^{f1,f2}$ for each possible feature context pairing $\{f_1, f_2\} \in F$. The Nearest Neighbour distance metric specified is the Bhattacharyya coefficient. The nearest neighbour distance metric for feature context pair $\{f_1, f_2\}$ is thus defined as:

$$B(Y_n, Y_m) = \sum_h \sqrt{Y(h)_n^{f1,f2} Y(h)_m^{f1,f2}} \tag{4.8}$$

Where we sum over all histogram bins h for feature dimension f_1. Thus given a

feature vector for individual $n \in N$ at frame $t \in T$ we find the nearest neighbour m where $\{m \in N : n \neq m\}$.

$$NN(Y_n) = \{Y_m \in Y | \forall Y_p \in Y : B(Y_n, Y_m) \geq B(Y_n, Y_p)\} \tag{4.9}$$

The nearest neighbour equation specifies m the index of the least distant behaviour profile of n for feature context pair $\{f_1, f_2\}$ and B the resultant Bhattacharya coefficient. As the Bhattacharyya coefficient is a measure of similarity, scoring more similar distributions higher, the NN finds the greatest Bhattacharyya coefficient to distribution Y_n from the set of all distributions Y given the feature context pair $\{f_1, f_2\}$, we then recombine the independent feature context pairs to generate a single value for the abnormality coefficient $A(n,t)$ for person n, at frame t. The abnormality coefficient of behaviour at frame t for person n is the least supported feature pairing; the lowest similarity to the nearest neighbour:

$$A(n,t) = \arg\min_{f1,f2} B(Y_n^{f1,f2}, Y_m^{f1,f2}) \tag{4.10}$$

A consequence of segmenting subgroups is that an observation may be the only member of a context defined sub group. Ideally in operation an active learning methodology would be implemented to determine the normality of an observation in a new area of the behaviour space. However, in our application we chose to suspend judgement of new instances of behaviour, specifying that no evidence of an alarm is not an alarm. It would be equally valid to select the opposite, the effect of which would be to place a bias upon highlighting rare behaviour.

4.5.3 Anomaly Detection

Threshold μ upon $A(n,t)$ separates anomalies from normal observations and in effect represents the sensitivity of the system. If we seek to detect only anomalies then μ represents the expectation of abnormal behaviour in the sequence. For the end user μ represents a constant surveillance workload for the operator. Variable μ can be either set by the operator or defined empirically in an additional training phase. Anomalies $A(n,t)$ at frame t for person n are classified by:

$$A(n,t) = \delta(A(n,t)) = \begin{cases} 1, & A(n,t) < \mu \\ 0, & A(n,t) \geq \mu \end{cases} \tag{4.11}$$

Based upon the assumption that there is dependence between the behaviour of individuals within the same social group we utilise the social contextual information in an additional two ways. Firstly we ensure that the behaviour of each individual is only analysed in reference to people external to their social group. Thus a behaviourally homogeneous group of individuals all acting abnormally cannot be self-justifying. We enforce this by removing the index of individuals from the same social group from the nearest neighbour calculation for individuals in that group. Secondly, social information enables us to propagate the expectation of an anomaly through the entire social group. In this way each member of a social group at any given frame has the highest anomaly score for all individuals in that group. Thus if one individual in a group is behaving abnormally all group members are equally as abnormal. We do not implement any post process alarm filtering. We justify the exclusion of this process as it may obscure the change in accuracy resulting from the inclusion and exclusion of contextual information.

4.6 Experiment

We wish to evaluate whether social and scene region contextual knowledge improves the detection of behavioural anomalies and permits the detection of subtle behavioural anomalies. We now detail the results of an anomaly detection experiment on the PETS 2007 dataset with the inclusion and exclusion of contextual information. Furthermore we test against a state of the art behaviour anomaly detection system which is itself designed to detect subtle anomalies.

The publicly available PETS 2007 dataset [2] offers a source of multi camera real world surveillance footage. The datasets consist of 8 sequences each captured from 4 different viewpoints. We consider the PETS 2007 data to be a crowded scene. The data contains a total of 573 individuals over 11902 frames, averaging 24 people in the scene at any given frame in a space measuring 16.2 meters by 7.2 meters. Behavioural anomalies in this dataset are characterised by strong motion abnormalities such as a group running across part of the scene, or subtle anomalies such as a single individual standing still in a busy area, or a group loitering amongst a crowd. We specifically chose this data due to its behavioural complexity for anomaly detection. The second dataset selected is the Oxford dataset. The Oxford data contains 430 tracked pedestrians over 4500 frames. There are an average of 15 individuals in any given frame, with a minimum of 5 and a maximum of 29. We consider this data as sparsely populated. The trajectory motion in the Oxford data is far more structured; the vast majority of individuals travel at walking pace in one of two directions. We select the second dataset, the Oxford data, to test our social context approach for failure modes. In the Oxford data the trajectories of socially unconnected pedestrians are often very similar, and often close in proximity - giving the appearance of social connectivity. We expect this will produce false positive social context information. We evaluate upon 3 non-sequential videos from the PETS 2007 selected due to the ground truth behaviour abnormalities present. PETS Scene 02 consists of 4500 images, Scene 04 is 3500 images long, and Scene 07 is 3000 images in length. All three are imaged at 25fps. The single scene from the Oxford dataset is captured at 25fps and 4500 frames in length. each sequence is treated individually. We apply the tracking procedure outlined earlier upon the jpeg the format images with no other pre-processing.

4.6.1 Scene Segmentation

We found well defined regions for the idle, divergence and traffic region in the PETS data which fit with the intuitive interpretation of the scene. For clarity we illustrate the scene segmentation, see Figure 4.4. The Oxford data held well defined areas for the traffic region and the divergence region. However the idle region hardly featured. This finding fits with the highly structured nature of the Oxford data in which there are very few stationary tracks. As our approach is data driven, scene regions are defined by virtue of being a tool for segmenting the behaviour space rather than fitting an intuitive interpretation of scene regions.

4.6.2 Social Context

We test the social context classification against an independently constructed ground truth for social connections. The training data (PETS 2006) consisted of 28 people with 14 true positive unique social connections between them of varying strength. The test data (PETS 2007) contains 152 tracked individuals, 44 social connections.

Classifying social connections in the PETS 2007 data using parameters trained in the PETS 2006 data achieved a True Positive Rate (TPR) of 0.92 and a False Positive Rate (FPR) of 0.092, see Figure 4.3 (a). There are a greater number of false positive social connections in the Oxford data. The optimal result found 0.412 TPR and 0.0149 FPR. However beyond this true positive rate the false positives escalated greatly.

4.6.3 Anomaly Detection

To demonstrate the impact context information has upon anomaly detection we determine the accuracy in four states: no contextual information, only scene context, only social context and with both types of contextual information. A comparison is made of the TPR and FPR, for detection of groundtruth anomalies. See Table 1 for a full list of anomalies. For examples of subtle anomaly detection see Figure 4.5. The anomaly ground truth reveals 12 behavioural anomalies in the PETS 2007, and 3 anomalies over 4500 frames in the Oxford data. In both the PETS and Oxford data we vary the μ threshold from 0 to 1 in small increments to adjust the system's sensitivity to unlikely observations. Figure 4.6 (a) (b) and (c) demonstrates the anomaly detection success in the PETS 2007 dataset. Figure 4.7 illustrates the results on the Oxford data.

4.7 Evaluation

The final TPR and FPR classification results with the inclusion of both types of context are affected by three factors above the no-context baseline. Firstly, the inclusion of scene context, the inclusion of social context, and impact of propagating anomalies through a social group and denying self-justifying social groups. In the three PETS-2007 datasets we observe that the addition of scene context improves the TPR over FPR detection of anomalies over all datasets in comparison to the no-context baseline. This is most significantly observed in Scene 04, Figure 4.6 (c). The inclusion of social context alone into the PETS-2007 data demonstrates a reduction in anomaly detection capacity in Scene 02, Figure 4.6 (c). PETS-2007 Scene 02 shows only a minor improvement. The significant result is that with the inclusion of both social context and scene context the TPR is improved above the TPR of scene context inclusion alone. This is due to the inclusion of the capability introduced by the social context to deny self-justifying groups and propagate anomalies within social groups. Particularly in PETS Scene 04, we observe that by propagating low likelihood scores throughout the group the bulk of true positive anomalies are discovered earlier, reducing the FPR from 0.2 to 0.03, see Figure 4.6 (c). The overall classification score with both social and scene context for all PETS-2007 data is shown in Figure 4.8. We recorded a drop in the false positive rate of 0.13 for the optimal classification rate of 0.78 when applying the social and scene context.

In the Oxford data set the use of context information does not appear to raise the ability to detect anomalies significantly. We believe this to be due to the highly structured simple nature of the Oxford data. There is in effect very little contextual information to leverage our method upon. The false positive social connections in the Oxford data has not adversely affected use of social context, however, the inclusion of denying self-justifying groups, and propagating anomalies through social groups has a notable negative impact. The impact of denying self-justifying groups in the presence of false positive social groups is to remove potential training data, thus

Table 4.1: The behavioural anomalies in PETS 2007 (3 sequences) and Oxford Data. (1), (2) and (3) occur due to a group standing on the left of the scene looking around and suddenly dispersing in different directions. Anomalies (4) and (5) occur due to two individuals entering the scene, turning a corner and then suddenly turning around and leaving in the same place they entered. (6) is a known ground truth behavioural anomaly. One of the participants in the PETS 2007 experiment purposefully loiters in a busy scene. (6), (7) and (8) are all members of a small group of 3 running through the scene, from the top to the bottom of the scene. (9), (10),(11), and (12) are four more instances of known ground truth anomalies. Two individuals purposefully loiter in the scene whilst another two suspiciously switch baggage. In the Oxford data, anomaly (13) is due to the unique behaviour of the individual interacting with a bin in the scene. Anomaly (14) captures an individual entering the scene at the bottom and loitering in the middle. Anomaly (15) captures a women meandering slowly through the scene.

PETS 2007 (Scene s00)	Id	Start	End
Unusual group behaviour	1	1	2656
Unusual group behaviour	2	1	2419
Unusual group behaviour	3	1	2714
Abrupt you turn in busy area	4	2627	2928
Abrupt you turn in busy area	5	2604	2928
PETS 2007 (Scene s02)	Id	Start	End
ground-truth loitering	6	160	4497
PETS 2007 (Scene s04)	Id	Start	End
Running through scene	6	109	275
Running through scene	7	130	290
Running through scene	8	148	322
Bag swap, unusual motion	9	1	3496
Bag swap, unusual motion	10	1	3496
ground-truth loitering	11	1	2596
ground-truth loitering	12	497	1726
Oxford Data	Id	Start	End
Motion + interaction with scene	13	3554	4349
Loitering	14	3867	4500
Abnormally slow movement	15	2382	3454

increasing the probability of false positive anomaly alarms. We observe this failure mode in the Oxford data, see Figure 4.7 which reflects our original prediction that our social model, geared towards crowds, would present a failure mode in the highly structured motion of Oxford data. To further test our approach we applied our context aware algorithm to maritime AIS shipping data in Southampton Harbour. The social context depicted mutual dependencies such as tugs pulling ships and convoy behaviour. Scene context was directly comparable. We achieved a true positive anomaly detection rate of 0.98 with a false positive rate of 0.17 over 66 hours of data. However as the focus of our approach is computer vision we do not discuss the results further in this work.

In the PETS-2007 data anomalies such as loitering are subtle behavioural anomalies as the trajectories of these behaviours are very similar to a large number of legitimate behaviours in the scene, particular in the queuing areas. Because motion alone is not sufficient to define the behaviour as an anomaly we require extra contextual information to segment these subtle behaviours from the main body of data,

particularly the scene context. The output of our system is displayed in Figure 4.5. Images (a) through (c) show correct identification of anomalies. Image (a) shows an example of a context independent anomaly: running through the scene. Image (b) shows two examples of context dependent anomalies. The motion features pertaining to the anomaly are common within the entire scene, requiring scene context for them to be detected as anomalies.

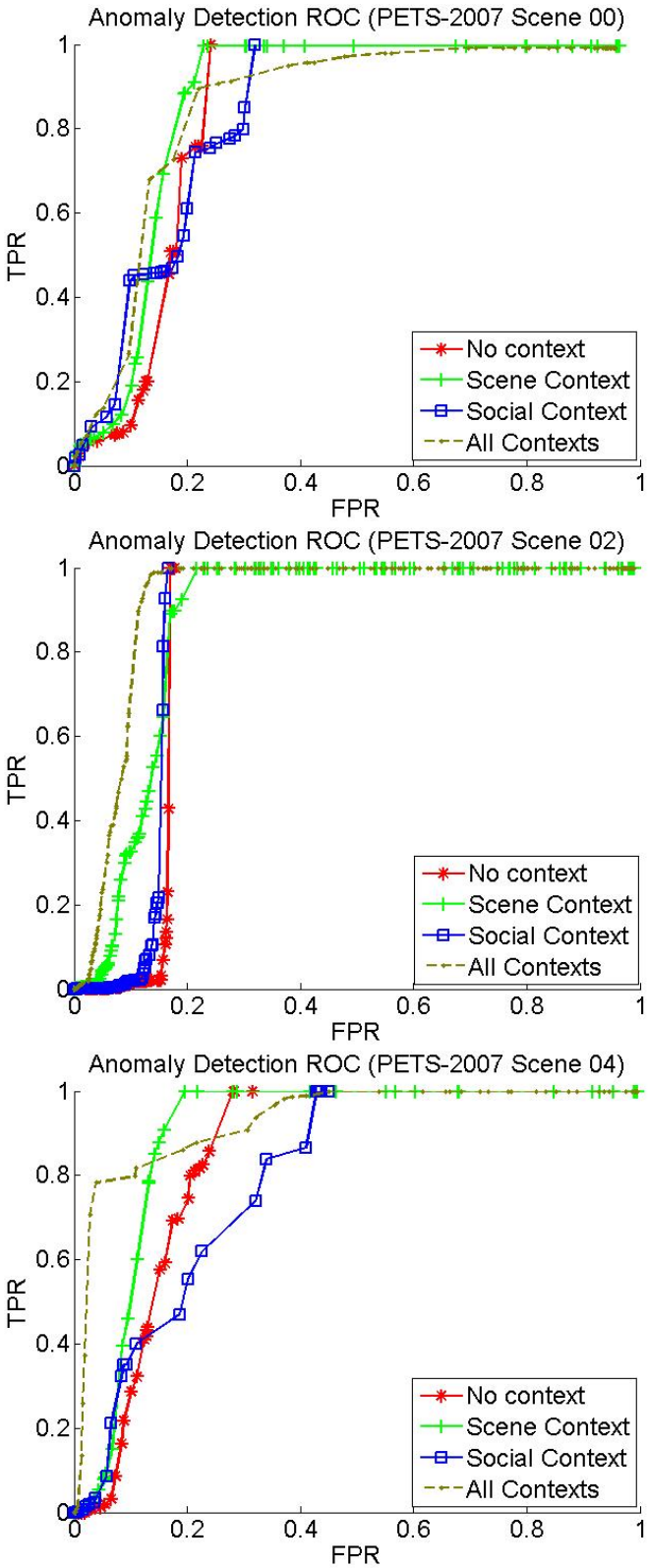
To see our anomaly detection system in reference to the state of the art we include an implementation of the Weakly Supervised Joint Topic Model (WSJTM) proposed and developed by T. Hospedales, Jian Li, Shaogang Gong and Tao Xiang. We select the WSJTM as it is designed specifically to detect *subtle* abnormal behaviour similar in style to our own work. Furthermore, it is based upon a different behaviour representation whilst its use of positional information makes it comparable to our scene contextual information. For a detailed account of this work see [37]. We use the code provided by the author to make the comparison. The results from our own and the WSJTM procedure can be seen in Figure 4.8. We find that the WSJTM outperforms our method at low TPR and FPR rates. However the results sharply fall off as it is incapable of segmenting a range of anomalies from the challenging PETS-2007 data. The WSJTM is capable of finding gross motion anomalies better than our method however it fails to detect subtle anomalies such as loitering. We observe that our method achieves a better overall TPR over FPR.

4.8 Conclusion

We successfully demonstrated the capability to detect anomalies based upon contextual information and target trajectories in two scenes, presenting distinctly different behavioural environments. The application of social context provides an improvement in anomaly detection in the crowded PETS-2007 data. However, failure of the social model can result in a negative impact upon anomaly detection, as witnessed in the Oxford dataset. We found that our context aware method performs significantly better than the equivalent method without contextual information; reducing the false positive rate from 0.2 to 0.03. We show an overall true positive classification rate of 0.78 over 0.19 false positives on the PETS-2007 data, a reduction in the false positive rate of 0.13 due to the inclusion of contextual information. We conclude that in a crowded scene the application of social context to prevent self-justifying groups and propagate anomalies is highly relevant. Scene context uniformly improved the detection of anomalies in both datasets, and provided the ability to detect subtle context dependent anomalies. The metric for comparing behaviours in this work can be interchanged with other state of the art methods; the implication being that contextual information, particularly scene regions, could be complimentary used with other anomaly detection systems revealing subtle anomalies that otherwise may be missed. Specifically, the novel contributions we make in this chapter are as follows:

- A novel method of acquiring scene structure information in surveillance
- The development of a novel social group classification algorthm using mutual information
- The demonstration that social and scene contextual information can improve the detection of human behaviour anomalies. Further validating the growing trend in automatic scene understanding.

We have now set out the fundamental validation that two sources of contextual information can be used to enhance human behaviour anomaly detection. This validation spurs on our research to further develop the contextual states to fully utilise the features available by incorporating in visual attention, see Chapter 5. The current work demonstrated the power of the contextual features within a simplistic anomaly detection framework where anomalies are characterised within the feature-space. Our later work examines the benefit of changing the basis from the current observable feature space to separate behaviour space; where behaviour can be encoded in a more efficient representation, becoming more invariant to expected behavioural variation, whilst retaining characteristic attributes that define the behaviour, see Chapter 6.



(a) (b) (c)

Figure 4.6: ROC charts for Anomaly Detection classification, with a comparison of different contextual setups. (a) shows the results from PETS-2007 Scene 00, (b) from PETS-2007 Scene 02, and (c) from PETS-2007 Scene 04.

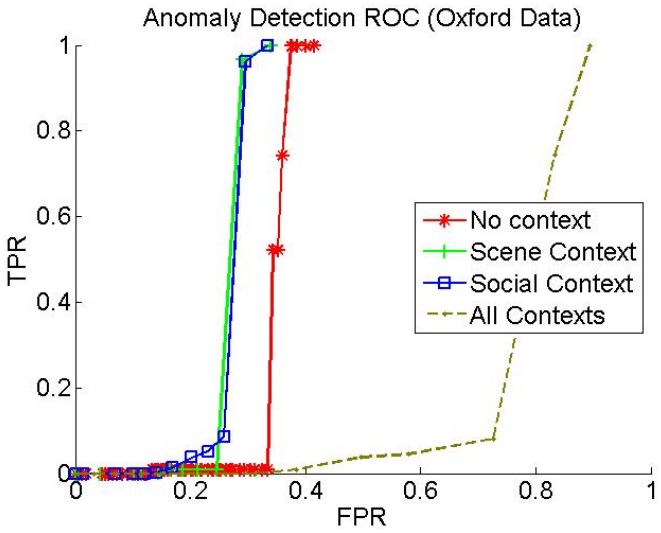


Figure 4.7: The anomaly detection results on the Oxford Dataset. we test upon the Oxford data to test for a failure mode in the social model.

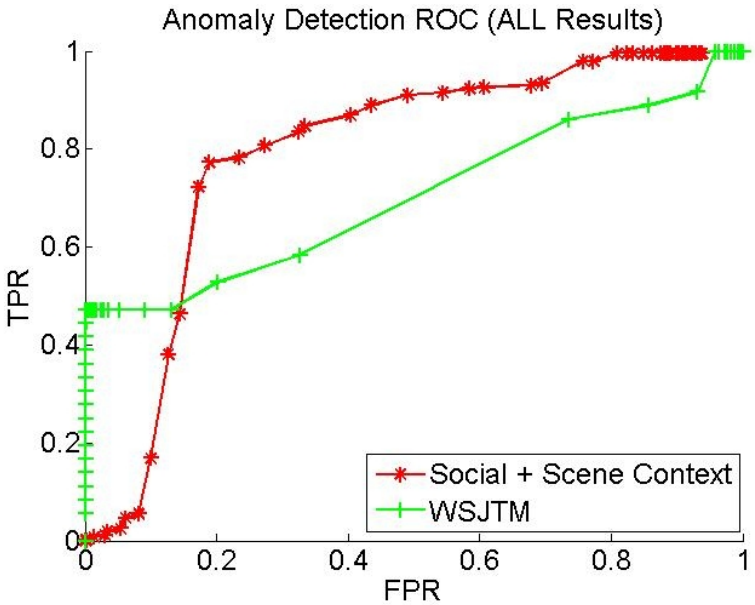


Figure 4.8: A comparison between the Weakly Supervised Joint Topic model and our context aware method on the challenging PETS-2007 dataset. We trained and tested against all PETS-2007 data for both datasets.

Chapter 5

Social and Scene Modelling with Visual Attention

In this chapter we describe the development of an improved social and scene modelling method which builds upon our previous work, see Chapter 4 and state of the art techniques 2. The aim of this work is to introduce the additional extractable feature of head pose, and the derived feature of visual attention into our social context and scene modelling work. Additionally we address the fundamentals of our previous algorithm, which proved the principle of using contextual information to overcome some weaknesses and develop a more principled approach. We find that we can classify human social groups in surveillance at a higher accuracy with visual attention. Additionally an intuitive contextual model of the scene is developed which incorporates the head pose feature. The research in this chapter addresses research objectives 5.2, 6.1 and 6.2; see section 2.7.

The social context work in this chapter was published in the following; in IEEE Signal Processing Letters [8], in Computer Vision and Pattern Recognition workshop 2014 [54], and in the Sensor Signal Processing for Defence conference 2014 [7].

5.1 Introduction

Human behaviour analysis has presented a challenging problem in autonomous surveillance due to the variety, subtlety, and obscurity of behavioural expression. In any non-trivial environment the interpretation of behaviour is often subject to the problem of heterogeneous behaviour grouping. This occurs when multiple classes of behaviour are indistinguishable or incorrectly classified due to limitations of the behaviour representation, or limitations on the extractable features. This is the case particularly in unsupervised methods; observations of normal behaviour draw from multiple behavioural classes with no obvious segmentation, and often a sparsity of examples prevents the learning of class boundaries in feature space resulting in the masking of outlier behaviours. Contextual information can be exploited to break heterogeneous behaviour classes into homogeneous classes [55]. The contextual information represents universally applicable a priori expectations for the domain. Scene segmentation, social clustering, and temporal segmentation are three common examples of contextual information which enhance the interpretation of behaviour. Estimating social connectivity between individuals is a contextual feature gaining popularity in recent work [70] [88] [24] [9] [64]. Social connectivity and grouping is

used to improve tracking [70] and behaviour analysis. It has been shown that with an understanding of the social context surrounding human behaviour in surveillance it is possible to better interpret observed events and detect abnormal behaviour [55] [37]. Using the social behaviour feature is particularly relevant in crowded environments in which the motion of an individual is more constrained and social dependencies are more salient against the entropic crowd motion.

Previous research has shown that context aware human behaviour anomaly detection based upon motion features alone is capable of detecting subtle abnormal behaviours [55]. We now progress this line of research by integrating the non-motion feature of visual attention into the behaviour representation. The intuition is that whilst motion features carry ambiguity in their behavioural implication due to external environmental influences upon the individuals motion the visual attention of an individual is free from the same environmental influences and may betray behavioural intents or interactions that the motion is incapable of displaying. Additionally there is a limitation of expression motion alone can provide; visual attention will provide additional information to characterise the behaviour. Furthermore, whilst motion features give an indication of past to current interaction with the environment, the visual attention of an individual may indicate the future intention of the individual. Although we do not test this explicitly; our focus is upon harnessing the visual attention profile information to better separate outlier abnormal behaviour from normal behaviour.

Our method advances existing social grouping methods which have focused primarily on motion features. We first review the background literature related to social grouping and estimation. To estimate social groupings Ge et al. uses a proximity and velocity metric to associate individuals into pairs, iteratively adding additional individuals to groups using the Hausdorff distance as a measure of closeness [32]. Yu et al. implements a graph cuts-based system which uses the feature of proximity alone [88]. However, modelling social groups by positional information alone is prone to finding false social connections when individuals are within close proximity due to environmental influences such as queuing. Oliver et al. uses a Coupled HMM to construct a-priori models of group events such as Follow-reach-walk together, or Approach-meet-go separately [66]. Certain actions are declared group activities and thus groups can be constructed from individuals via mutual engagement in a grouping action. However, a more recent development in automatic social grouping seeks to model social interaction using the visual interest of the tracked individuals. The use of an individual's visual attention is significant as it uses a rich feature which indicates the intention of the individual. Robertson and Reid utilize gaze direction, also referred to as head pose direction, in order to determine whether individuals are within each other's field of view [74]. Farenzena et al use an estimation of the visual focus of attention of a person as a cue to indicate social interaction [9]. Head pose is quantized into 4 different locations at each frame, and a predefined set of spatial and visual criteria determines if the conditions for a social interaction are met at each time step. A social exchange is then defined as lasting a given duration (10 seconds). In our work we bring together the motion-based social paradigm with the benefit of visual information as it is demonstrated by [74] [9].

The use of visual attention departs from the focus upon motion information. We hypothesise that this approach will bring strength to human behaviour anomaly detection as it is complementary of motion and does not suffer from the same ambiguities as motion. Several factors impact the motion of an individual; the scene (doorways, high traffic lanes, queuing areas), environmental interactions (shops, cash

machines, bins), social motion dependencies (groups, following, meeting/departing events). These features can influence the motion of an individual giving the impression of abnormal motion if the context in not properly understood. However, visual attention is not similarly constrained by environmental context. Visual attention is often guided by environmental interest points [10], so it requires modelling in context, however visual attention is guided and restricted by different factors than motion, providing an additional source of behavioural information. See Chapter 3 for details as to how head orientation is extracted.

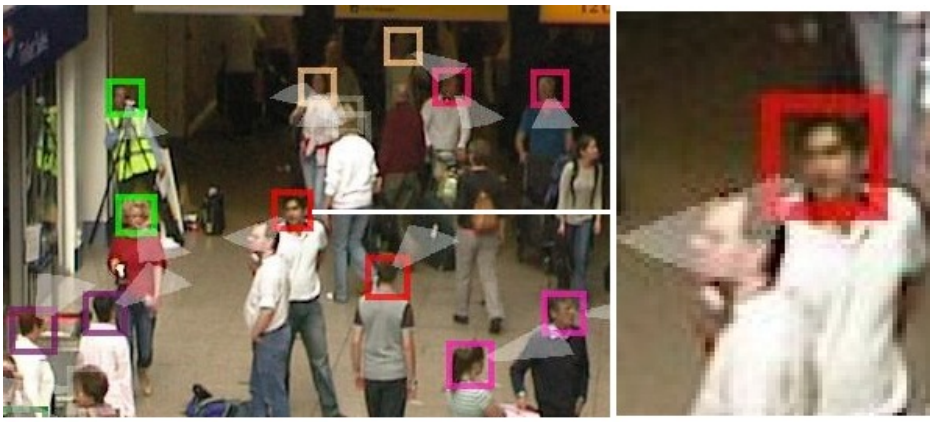
5.2 Social Grouping Using Visual Attention

Estimation of social connectivity between individuals is a contextual feature gaining popularity in recent work. Social connectivity and grouping is used to improve tracking [70] and behaviour analysis [55]. It has been shown that with an understanding of the social context surrounding human behaviour in surveillance it is possible to better interpret observed events and detect abnormal behaviour [55] [37]. Using the social behaviour feature is particularly relevant in crowded environments in which the motion of an individual is more constrained and social dependencies are more salient against the entropic crowd motion. Our work focuses on the use of visual attention to better classify social connections in a semi-crowded surveillance scene. Motion information of individuals and of crowds is commonly used in automatic social grouping, however, the surveillance environment can exert influence upon trajectories by channelling people, presenting queuing or waiting areas, or containing objects to interact with. These motions are ambiguous with intentional motion from social connections and as such obscure any trivial definition of social connectivity. In this work we extract a further feature; head pose, and derive from it the additional feature of visual attention, and demonstrate that visual attention can be used to better identify social grouping in crowded environments. The visual attention of an individual provides an additive feature which supplements the motion-based similarity used in the state of the art. The visual attention feature is not impacted in exactly the same way as the motion features are by the scene, as it is not influenced in the same way by the scene constraints mentioned earlier.

With this section of our research we aim to verify the hypothesis that pedestrian visual attention can be used to compliment motion-based social group estimation. To verify our hypothesis we will implement our hybrid motion-visual attention system demonstrating better social grouping in a variety of different surveillance datasets. Comparison will be made against a hand labeled social group ground truth, assessing the efficacy of our visual attention and motion against motion alone.

5.2.1 Initial Hypothesis Validation

Our visual interest social grouping is based on the hypothesis that socially connected people act as a source of visual interest for each other. This hypothesis makes the implicit assumption that the gazing patterns of socially connected persons differs from those that are unconnected. The validation of the underlying hypothesis was performed in two steps. In the first step, pedestrians were segmented into two groups: those with social connections, and those without. This segmentation was performed by hand. Once segmented, we calculated the deviation between travel direction and head pose for each pedestrian for each frame of video. Travel direction was calculated using each persons smoothed velocity over a 15 frame window using



(a) (b)

Figure 5.1: Image (a) illustrates an example from the PETS 2007 dataset [2] of our tracking output, the social groups (designated by coloured bounding boxes) and the extracted head pose estimates (illustrated by field of view cones). image (b) is a zoomed subsection showing the visual field of view of a person in the image.

the head centroids provided by Benfold [12] and our own tracks on the PETS 2007 data [2]. In our initial validation we removed false-positive head detections and used hand-labeled head pose rather than utilizing algorithmic solutions. Formally, denote a persons velocity direction at frame t as θ_t^D and their head pose as θ_t^G. The head pose velocity deviation can then be calculated as the absolute error $\epsilon_t = |\theta_t^D - \theta_t^G|$. The mean and variance of the deviations was then extracted for the two pedestrian groups (socially connected and unconnected) upon which further analysis was performed.

Validation Results The analysis of gazing patterns was performed on 3 datasets: the Benfold dataset [12], the Caviar dataset [1] and the PETS 2007 dataset [2]. In each case the pedestrian detection and tracking information provided by each dataset was used. Where not supplied, additional ground-truth head pose labels were added. Statistics were extracted for 37 tracks from the caviar dataset, 372 tracks from the PETS dataset, and 170 tracks from the Benfold dataset. Figure 5.1 shows example frames from PETS scene 4 highlighting socially connected and unconnected persons. We illustrate in Figure 5.2 the extracted distributions from all datasets. One can see from the figure that for the Benfold dataset, there is little difference between the gazing patterns of the two groups. However, on the caviar dataset two distinct distributions are observed, as is also the case with the PETS dataset. For each dataset, performing the χ^2 variance test between the socially connected and unconnected deviations with a p-value of 0.05 shows that in all three datasets, the differences between the deviations for socially connected and unconnected persons are statistically significant.

To partially validate our assumption that socially connected individuals are a source of visual focus for each other, we analysed the null hypothesis that socially connected and unconnected persons have the same gazing patterns. Our analysis of deviations between travel direction and head pose direction showed evidence that gazing patterns do differ between socially connected and unconnected persons. However, the degree of separation between distributions varied for each dataset, identifying the need for the weighting factor to be used when using head orientation data for determining social connectivity. For all datasets the differences between

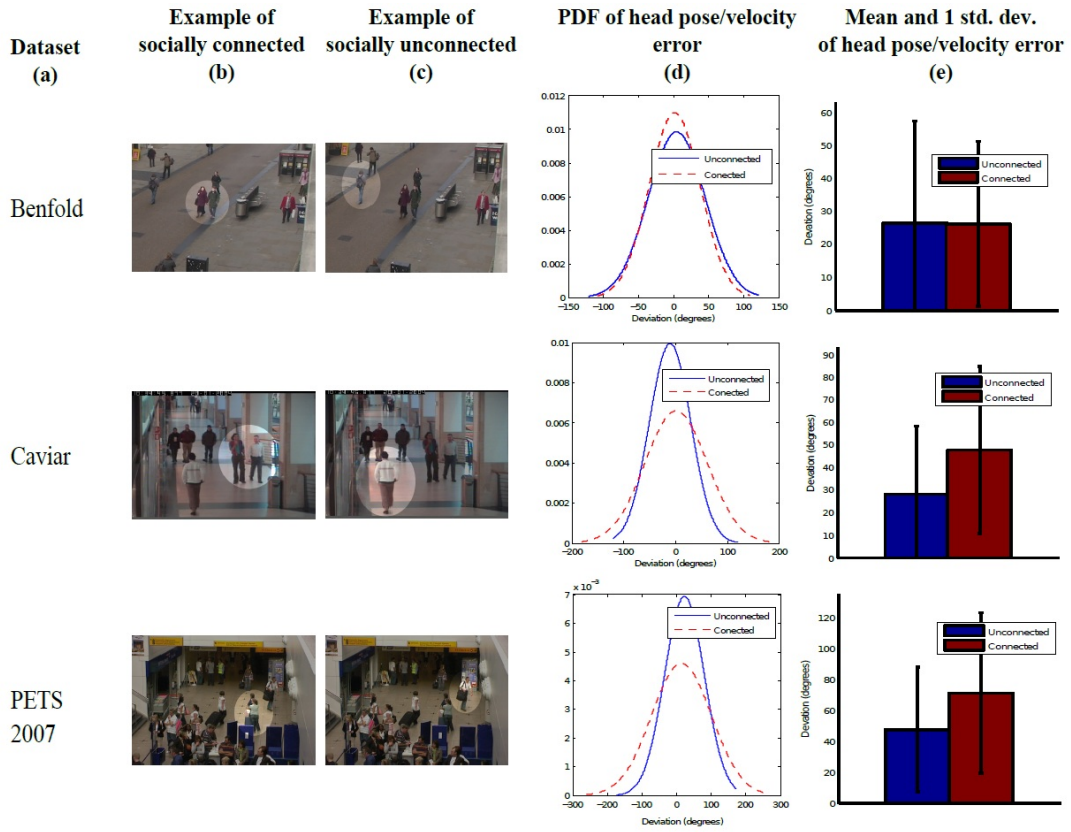
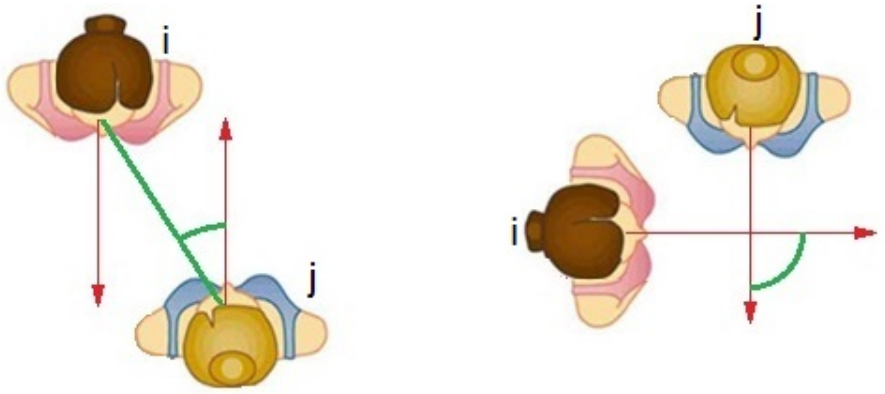


Figure 5.2: Example frames and extracted head pose velocity deviation (error) statistics extracted from three datasets. Column (d) shows the normal distributions for socially connected (red) and unconnected (blue) persons. Column (e) shows the mean and standard deviation of socially connected (red) and unconnected (blue) persons.

the two groups were statistically significant giving support for our assumption and leading us to reject the null hypothesis.

5.2.2 Social Modelling using Visual Attention

Our previous motion-based social grouping, see section 4.4, is grounded upon the premise that shared trajectory information implies a social dependence between two individuals. The principles of the social force model are such that socially connected individuals are more likely to move together, and thus display more similar trajectory information, and socially independent people feel a force of repulsion and are more likely to *avoid* moving similarly and avoid close proximity. The more entropic the underlying motion of the crowd the more salient similar social trajectories will be. The grouping method finds social similarity within the features of direction, speed, proximity, and temporal overlap. Each feature is weighted based upon a one off training phase, such that proximity and temporal overlap have more dominance in the overall metric than direction and speed, which were found to be less important. The similarity of direction and speed are measured using the mutual information measure. The proximity and temporal overlap similarity are measured by euclidean distance. Once the similarity for each feature has been measured the four features are combined to a single similarity measure, 5.4. Each tracked object has a similarity to every other, populating a social pairing likelihood table.



(a) (b)

Figure 5.3: Illustration of mutual visual attention (a) and visual correlation (b). In both cases the red line represents the head pose direction for person i; variable θ_{it}^{G}. In (a) the green curve is the the direction from person i to j, the angle of mutual attention; θ_{ijt}. For image (b) red lines represent each individual head pose angle and the green curve represents θ_{jt}^{G} the angle of visual correlation difference for person i and j.

To verify our hypothesis that visual interest can be used to enhance the existing motion-based social grouping we incorporate head pose direction and subsequently visual interest into the social grouping model. The distinction between head pose direction and visual interest is as follows; head pose direction is the raw angle in which the person is looking, usually indicated by head pose in our data, and visual interest is a distribution over possible regions of interest. In our case, we extract head pose estimates in order to estimate visual interest using knowledge of interest points and characteristics of how interest drops at the periphery of vision and with distance, permitting an estimation of the visual interest any given person has in their environment.

We theorize that there are two ways socially connected individuals can demonstrate social interaction through head pose direction; correlated direction of visual interest, and looking at each other. The former occurs in cases when two individuals are actively looking at the *same thing*, which requires communication to coordinate, however it could be coincidental when an event or object has drawn both of their attention. The latter event, when two individuals are looking at *each other*, implies that they are the object of attention for each other. It is at least unusual for two socially unconnected individuals to look at each for a prolonged period of time. Following from this reasoning, there are two events we wish to measure. These are, how similar the head pose direction of two individuals are and the amount of time the head pose is directed towards each other. Figure 5.3 illustrates the two examples of mutual visual attention (a) and visual correlation (b). In both cases the red line represents the head pose direction for person i; variable θ_{it}^{G}. In (a) the green curve is the the direction from person i to j, the angle of mutual attention; θ_{ijt}. For image (b) red lines represent each individual head pose angle and the green curve represents θ_{jt}^{G} the angle of visual correlation difference for person i and j. We wish to exclude cases where two individuals are looking in the same direction due to walking in that direction. The work of Benfold [12] showed that pedestrians spend the majority of their time looking in the direction of travel. To avoid highly scoring correlated head pose due to two people looking in the same direction of travel we

introduce a weight which represents our confidence that the direction of head pose is due to visual interest other than direction of travel. The weighting is greater for those with a head pose direction off their current direction of travel. The visual correlation weight coefficient is given by:

$$\omega_{ijt} = |\theta^G_{it} - \theta^D_{it}||\theta^G_{jt} - \theta^D_{jt}| \tag{5.1}$$

Where θ^G_{it} is the gazing direction for person i at frame t, and θ^D_{it} is the direction of travel. Thus we score people with attention towards either side stronger than those who are looking in the direction of travel, dropping linearly. This is justified on the assumption that a social source of attention is more likely when not looking in the direction of travel; backed up by the preliminary hypothesis verification. If there is no current direction of travel then this weight is always 1. Similarly, we introduce a weighting for visual mutual attention. We weight the measure of visual interest between two individuals by proximity. The further away someone is the less confident we are they are a social focus of attention. The mutual visual attention weight is given by:

$$\lambda_{ijt} = 1 - \frac{\sqrt{x^2_{ijt} + y^2_{ijt}}}{X} \tag{5.2}$$

Where x_{ijt} and y_{ijt} is the x and y distance between person i and person j at frame t and X is the width of the scene; the maximal distance between two people. Thus we model the probability of interest between person i and person j as falling linearly with distance. We then define the total Visual Interest feature Λ_{ijt} between person i and j at any given frame as the product of two Gaussian distributions encompassing the visual correlation variance σ_λ and the visual mutual attention variance σ_ω predefined as $\pi/4$.

$$\Lambda_{ijt} = \frac{1}{\sigma_\lambda \sigma_\omega 4\pi^2} e^{-\frac{|\theta_{ijt} - \theta^G_{it}|^2}{2\sigma^2_\lambda} - \frac{|\theta^G_{it} - \theta^G_{jt}|^2}{2\sigma^2_\lambda}} \tag{5.3}$$

Where θ_{ijt} is the direction from person i to person j at time t. We next incorporate the visual interest into our system as another feature in the existing social similarity metric. We measure the visual interest similarity between each potential socially connected individuals and include this with a weighting of 1 into the social similarity metric. Thus for any two people the features that determine grouping likelihood in the social pairing table are; proximity, temporal overlap, direction, speed, and visual interest. The total social grouping strength between person i and person j for all frames is then given by:

$$\kappa_{ijt} = \frac{1}{T} \sum_t^T IV_{ijt} I\Theta_{ijt} \Delta P_{ijt} \tau_{ijt} \Lambda_{ijt} \tag{5.4}$$

$\tau_{ijt}, IV_{ijt}, I\Theta_{ijt}, \Delta P_{ijt}, \lambda_{ij}$ are the temporal overlap, mutual information for speed, mutual information for direction, proximity and visual interest difference between person i and j. Specific definitions for the motion features $\tau_{ijt}, IV_{ijt}, I\Theta_{ijt}, \Delta P_{ijt}$ are given in section 4.4 and Leach et al [56].

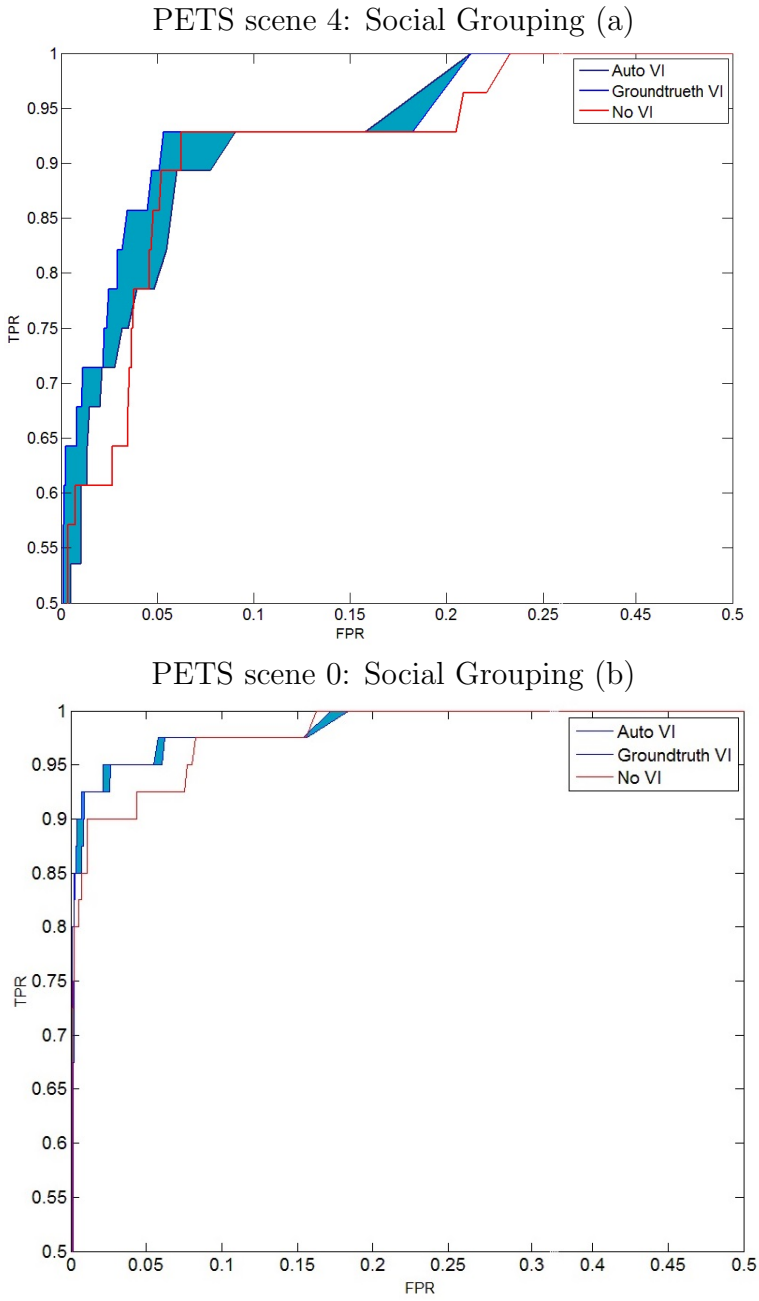


Figure 5.4: TPR and FPR from the pair connection likelihood matrix for the method without head pose information (red) and with head pose information (blue). The blue band illustrates the the results with groundtruth head pose direction as it degrades when automatic head pose direction is used. Image (a) shows the results for PETS scene 4 and image (b) for PETS scene 0.

5.3 Validation of Social Grouping

We wish to evaluate whether the use of visual focus of attention is indicative of social engagement, and if these features can be used to better classify social groups in multiple surveillance datasets. We evaluate the strength of the visual interest features by a comparison of the motion-based social grouping and the motion with visual interest social grouping. We test upon the publicly available PETS 2007 dataset [2] and the publicly available Oxford town centre data [12]. The PETS data offers a source of multi camera real world surveillance footage. The datasets consists

of 8 sequences each captured from 4 different viewpoints. We consider the PETS 2007 data to be a crowded scene. The data we use from this dataset contains a total of 372 individuals over 8000 frames, averaging 24 people in the scene at any given frame in a space measuring 16.2 meters by 7.2 meters. Social groups in this scene are characterized by small clusters of 2 - 4 people typically moving together or waiting together. The exception to this are four individuals who are actively engaging in abnormal loitering behaviour which separates them for relatively long periods of time. These individuals talk to each other at times in the scene and stand together at times, and as such are still considered to be socially connected. The Oxford data contains 430 tracked pedestrians over 4500 frames. There are an average of 15 individuals in any given frame, with a minimum of 5 and a maximum of 29. We consider this data as sparse to moderately populated. The trajectory motion in the Oxford data is far more structured; the vast majority of individuals travel at walking pace in one of two directions. In the Oxford data the trajectories of socially unconnected pedestrians are often very similar, and often close in proximity - giving the appearance of social connectivity. It is our prediction that the visual interest of pedestrians in this scene will be a relatively strong feature to detect social groups given the motion similarity of socially disconnected people. We evaluate upon 2 non-sequential videos from the PETS 2007. PETS Scene 00 consists of 4500 images, and Scene 04 is 3500 images long. both sequences are imaged at 25fps. The single scene from the Oxford dataset is captured at 25fps and 4500 frames in length. We apply the tracking procedure outlined earlier in section 3.4 upon the Jpeg formatted images with no other pre-processing.

5.3.1 Visual interest social grouping

We illustrate below the true positive rate TPR and FPR social group classification result for the three sequences we evaluated upon. In each case we ran the motion only social grouping method, the automatic visual interest and motion social grouping, and the groundtruth visual interest and motion social grouping. We use both groundtruth and automatic gazing direction estimates to illustrate the theory under ideal conditions, and to demonstrate the impact of noisy data. The output of our social grouping is a social connection likelihood matrix entailing the likelihood of each pair of individuals being socially connected, as detailed in the pair strength equation 5.4. All possible pairs of individuals have a probability of being connected. Applying a grouping strength threshold would thus define a grouping hypothesis stating a set of pairs. The grouping likelihood matrix implicitly contains many possible grouping hypothesis, each hypothesis characterized by a different grouping strength threshold. To find the true positive and false positive connections for different grouping thresholds we vary the grouping threshold from 0 to 1 in increments of 0.001; the hypothesis varies from 'no social connections' to 'everyone in one social group'. We find for the following optimal social grouping results by varying the connection threshold:

We find that in each dataset the inclusion of automatic head pose direction into the social grouping model improves the social grouping capabilities for the optimal threshold. For all thresholds we illustrate the improvement that the inclusion of visual attention provides in the social grouping efficacy Figures 5.5 5.4. The visual attention feature is a subtle and inherently noisy feature, and the motion only method achieves a result close to optimal, as such the improvements are only a small percent of the total value. For the Oxford data, we see a 5.6% improvement in true

Dataset	Auto TP/FP	GT TP/FP	Motion TP/FP
Oxford	0.90/0.07	**0.93/0.05**	0.88/0.07
PETS S4	0.89/0.06	**0.93/0.05**	0.89/0.06
PETS S0	0.93/0.02	**0.95/0.02**	0.92/0.04

Social Grouping Optimal Results

Table 5.1: We illustrate here the optimal social grouping result, selected from the ROC curves 5.5, 5.4. For the Oxford data, we see a 5.6% improvement in true positives and 28.5% reduction in false positives. For the PETS scene 4 data we see a 4.5% improvement in TPR and a 16.6% decrease in FPR. The PETS scene 0 data yields a 3.3% increase in TPR and a 50% decrease in FPR. Ground truth head pose scores highest for all three sequences and has joint or lowest FPR.

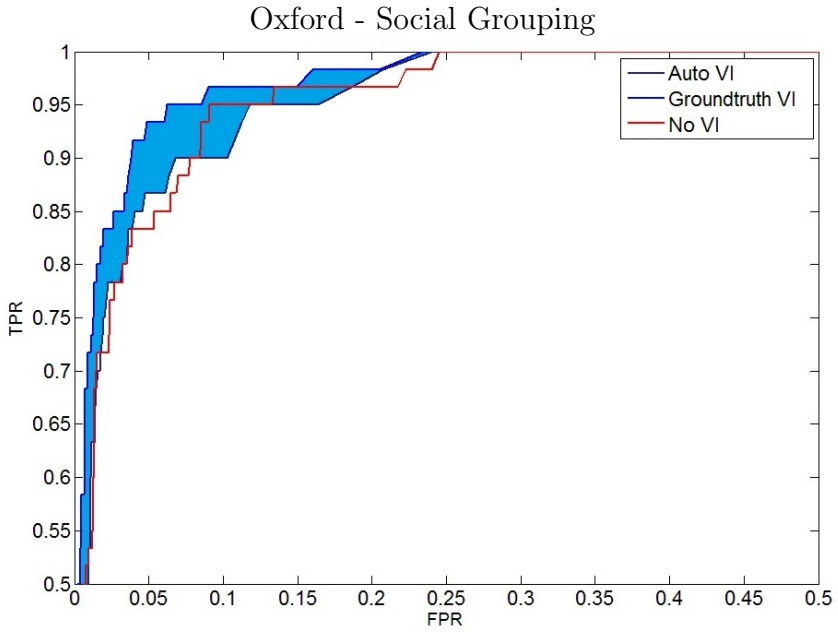


Figure 5.5: TPR and FPR from the pair connection likelihood matrix for the method without head pose information (red) and with head pose information (blue). The blue band illustrates the the results with groundtruth head pose direction as it degrades when automatic head pose direction is used.

positives and 28.5% reduction in false positives. For the PETS scene 4 data we see a 4.5% improvement in TPR and a 16.6% decrease in FPR. The PETS scene 0 data yields a 3.3% increase in TPR and a 50% decrease in FPR.

Figure 5.6 illustrates four examples of correct social grouping classification (Red, Pink, Blue, Dark Blue). Matching coloured bounding boxes signifies a social group classification by our algorithm. The group in red consists of three out of the four actors in the scene. The fourth is not included due to entering the scene significantly later. This grouping is particularly exemplary as the group starts off clustered but then splits up and loiters independently in the scene with high distance to one another. However, our system is capable of maintaining the social connection estimation between these people. The lighter blue grouping of the family in the bottom left of the image is an easy classification case as they share similar motion with a close proximity. Our tracker did not pick up the child in the group, thus

Figure 5.6: Illustration of 4 true positive social groupings (Red, Pink, Blue, Dark Blue). Matching coloured bounding boxes signifies a social group classification by our algorithm. The group in red consists of three out of the four actors in the scene. The fourth is not included due to entering the scene significantly later.

she is invisible to our algorithm. The stationary couple in pink bounding boxes are correctly classified, along side the two men in the top right of the scene, indicated with dark blue bounding boxes.

Figure 5.7 illustrates a further four social classifications. The two light green bounding boxes indicate a social connection between one of the actors in the scene and an official; both were seen communicating making this classification correct. The two dark blue bounding boxes, right hand side of image, capture two individuals sharing similar trajectories through the scene; their visual attention give a strong indication of social connectivity. The dark green bounding boxes indication two true positive connections between the closest two bounding boxes (man and women) and a false positive classification to the women in the queue. The stationary motion of all three gives little to distinguish true or false connection. It is possible in this case that it is the visual attention pattern that is responsible for the social connection, however this has not been explicitly tested for.

Figure 5.8 illustrates 4 social groupings (Red, Pink, Orange, Dark Green). The dark light green bounding boxes are the same as seen in Figure 5.7. The two pink bounding boxes indicate a social connection between two of a group of three. Therefore there is a false negative associated. Similarly we detect only two of three members of a family group, indicated by orange.

5.3.2 Discussion

Our results provide a strong indication that the inclusion of visual attention improves the capability of the motion-based social grouping in crowded human surveillance. We tested upon three video sequences; two PETS sequences considered challenging due to motion complexity, occlusion and crowding, and the Oxford data which is challenging due to a highly structured scene which masks salient social motion. We note that our system shows a susceptibility to head pose direction feature noise.

Figure 5.7: Illustration of 4 social groupings (Red, Green, Dark Green, Dark Blue). Matching coloured bounding boxes signifies a social group classification by our algorithm. The two light green bounding boxes indicate a social connection between one of the actors in the scene and an official; both were seen communicating making this classification correct.

An angular error of average 25 degrees can reduce the efficacy of visual attention and motion social grouping to below that of motion alone in the worst cases 5.5. However, the predominant result is an improvement when using automatic head pose direction above motion alone, and an even greater improvement when using ground truth head pose direction.

The power of the visual attention feature is that it is independent from the motion influences the environment presents, such as channelling people and queuing areas. The use of the visual attention feature is clearly additive to motion-based social grouping. There is however a computational cost to extracting head pose direction features from data. We computed head pose direction estimates as a batch process taking between 8 to 10 hours. However, Benfold [12] has demonstrated this process can be achieved at video rate when the feature space is sub-sampled, and still achieving good accuracy.

Our visual attention social grouping demonstrates, for the first time, the use of visual information in a generalized social grouping task, rather than used to detect specific or anecdotal social events. Our work demonstrates the applicability of visual information upon real world surveillance tasks, using a fully automated system. Our approach is most applicable to scenarios in which there is high motion similarity between social grouped people and un-grouped people, such as airports, stadiums, train stations, and busy town or city surveillance, particularly for use with automated human behaviour analysis.

5.4 Scene Modelling using visual Attention

We similarly wish to enhance the scene modelling with visual attention. The original scene model was validated using only motion information to segment the ground plane into 3 different types of region; traffic regions, idle regions, and conver-

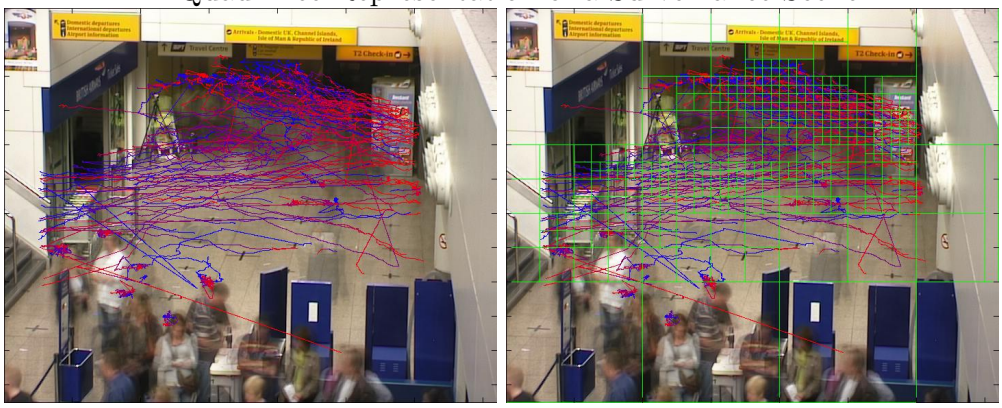
Figure 5.8: Illustration of 4 social groupings (Red, Pink, Orange, Dark Green). Matching coloured bounding boxes signifies a social group classification by our algorithm. The dark green bounding boxes are the same as seen in Figure 5.7. The two pink bounding boxes indicate a social connection between two of a group of three. Therefore there is a false negative associated.

gence/divergence regions. We make 3 fundamental changes in our methodology at this point. Firstly we introduce the concept of using a quad tree to segment the ground plane into atomic regions by how many unique track Ids have been through the area. Secondly, we match areas together based upon high similarity without defining in advance the possible region classes. Thirdly, we introduce visual attention into the calculation as an additive feature.

5.4.1 Quad Tree

We use a QT representation on the ground plane. The QT representation segments the ground plane into different sized squares, where each square contains the same number of tracked individuals. A Quad tree is similar to a binary tree with the major difference being that it has four nodes (one for each quadrant) instead of two. Each node, except for leaf nodes, points two four additional quadrants at the next lowest level of the tree. This method provides a spatial granularity that is dependent upon the amount of information present at each point in the image plane. We invoke the QT method in our work for three reasons. Firstly, the pixel size granularity of the image space is below the tracking error introduced to the system, and we thus do not wish to represent our environment spatially to this degree. The quad tree method grants a more meaningful coordinate system to represent position where the distance between points is dependent upon the information density between the two points. Secondly, there is a practical constraint on how fine a resolution the environment can be modelled at enforced by the memory space of the computer. If we model the environment, particularly the maritime environment, at a high resolution we hit this limitation, but modelling at a lower resolution sacrifices exactness in high density areas, thus a variable spatial resolution based upon information density is desirable. The third reason for using the QT representation is that it appeals to the notion of increasing interclass distance that we base our outlier detection behaviour analysis upon. We define later 6.5, that part of the measure of similarity between any two

A Quad Tree Representation of a Surveillance Scene



(a) (b)

Figure 5.9: A Quad Tree representation of the image space. The lowest nodes in the tree form the spatial quantisation on the ground plane. Each node is split to 4 smaller nodes if greater than 5 unique IDs travel through the node. Thus our coordinate system is a function of information density allowing an efficient encoding of the behaviour space. Image (a) shows all the trajectories over two minutes of data, where each trajectory is colour blue to red over time to show direction. Image (b) overlays the corresponding quad tree in green.

behavioural events is the proximity between the two events. However proximity is represented by two components; the quad tree distance scaled by node size, and the similarity of behaviour within the corresponding QT nodes the events come from. The former declares that distances in dense areas are exaggerated and distances in very sparse areas are contracted. The latter part of the metric ensures that behaviours have a low distance to other behaviours in similar regions, such as two queuing areas. The result is the effective warping of the scene; expanding and contracting areas of the scene based upon the density of tracks, and pulling similar areas closer whilst separating distinct areas. An illustration of this notion is provided in Figure, 6.2.

To implement the quad tree, we assume one high level node to cover the entire visual area, which is then split into 4 equal sized smaller nodes when the criteria for number of track inclusions is reached. For our method we split a node when 5 unique tracks are identified in a node. It is necessary to specify a minimum node size so that splitting does not persist Ad infinitum when a greater number of tracks than the splitting criteria exist on the same spot. The lowest nodes, leaf nodes, in the quad tree form the squares on the image plane and represent our coordinate system in Behaviour space. The benefit of such an approach is twofold. Firstly a quad tree representation of the image plane provides a basis for efficient spatial representation. We do not wish to accumulate large amounts of data for unpopulated areas when modelling behaviour, and by accumulating information at a spatial resolution based upon density we prevent over representation. This becomes more important in larger surveillance environments, particularly in the maritime domain. Secondly, the quad tree provides information uniformity meaning we scrutinise behaviour at a greater fidelity in more dense areas, taking advantage of the fact there is more training data for the location.

5.4.2 Region Grouping

Unlike our previous implementation of scene context, see section 4.3, we do not fully segment a scene at this stage. That is to say we do not enforce hard barriers between segmented regions as it introduces discontinuous statistics spatially. Instead we rely upon the behaviour metric defined later 6.5 to produce distance between behaviourally distinct spatial regions. Our behaviour metric defines distances between any point in behaviour space as a continuous function and thus we remove the impact of discontinuous boundaries within a heterogeneous dataset. A further benefit of not classifying regions with hard boundaries, as was implemented in our earlier work, is that to classify a region a definition had to be formed first. Which is to say the system required a priori definitions of all possible scene region classes. This entails that no distinction between regions can be used to create space between behaviour clusters in behaviour space unless the region has been predefined by an operator. This objectively limits region definitions to previously seen regions which can be expressed by the observable feature space, and subjectively limits region definitions to humanly intuitive scene regions. It is possible that there are spatial divides between behaviour clusters which are not intuitive to the humans perception of the scene, or are subtle enough that the distinction is overlooked. By taking a data driven approach and creating distances between behaviour clusters based upon statistical dissimilarity in the behaviour metric requires no previous definition of a spatial region, and no requirement for operator input. The weakness to this approach is that the system may become more prone to over-fitting. If spurious trends in the data arise a behavioural distinction may follow which does not truly represent the generic behavioural scene. This risk is somewhat mitigated by the fact our system is adaptive over longer periods of time.

We illustrate below the different features which compose the scene context aspect of the behaviour metric. Fundamentally driving the scene region context element of our behaviour analysis is QT node similarity measure. The similarity between any QT nodes is defined by a mixture of features; speed, direction, and visual attention. In fact each feature uses the magnitude of the distribution of the feature in a quad tree node as well as the entropy of the feature. By this definition behaviour observations are closer related to other behaviour observations in QT nodes with a similar distribution of speed, direction and visual attention. This is so that when seeking to justify a behaviour profile we look for evidence only within a similar spatial context. The justification for this approach is extensively verified in the chapter on context aware anomaly detection 4. In brief the reasoning is because in some cases a behaviour is abnormal not because of the pattern of motion itself but because of the location of the pattern. For example, a person walking swiftly through an area full of stationary people is more abnormal than a person walking swiftly in a high traffic region. Equally a person standing motionless is normal only when in the context of a queuing area. Thus when seeking to justify the stationary motion is it important to look for similar examples only within the same or similar region context. The similarity between QT nodes for each feature is illustrated below. The colour intensity map, Figure 5.10, given in (e) ranges from cool to warm where warm indicates higher values in the normalised feature map. Image (b), feature energy, measures the spread of the distribution, and image (c) illustrates the entropy, or how flat the distribution is; where a two peak distribution has low entropy but high energy.

We observe that the distribution of speed is well structured in the scene, as can be seen from Figure 5.10 image (a). Running horizontally is a band of higher mean

speeds (a). Conversely the rest of the scene, with the exception of a small number of nodes, have consistent low velocity motion, as shown by the low feature magnitude image (a) and the low entropy of the feature distribution, image (b). The energy and entropy maps illustrate the shape of the feature distribution. Entropy measures how flat the distribution is, whereas energy effectively measures the standard deviation of the distribution, thus capturing two separate characteristics of the distribution. The band of high velocity shown in image (a) has high distribution entropy and energy indicating that whilst there are high velocities there is also a mix of high and low velocities. Image (d) shows the similarity score between neighbouring QT nodes. The purpose of image (d) is to emphasise structure in the scene. Only the similarity to between neighbours is shown, so structure of depth 2 nodes or more is not apparent from this plot.

5.4.3 Region Similarity Definition

The cross node region similarity score, which measures the similarity between QT nodes on the ground plane is a linear summation of energy, entropy, and magnitude of the feature distributions. In effect we measure the similarity of the shape of distributions of features observed in the QT nodes, where the features are the distributions of speed, direction and visual attention of tracked pedestrians travelling through the node. We define the energy of a distribution X as:

$$E(X) = \sum_i X(i)\sqrt{X(i) - \hat{X}} \tag{5.5}$$

Making the energy $E(X)$ of distribution X a measure of the standard deviation of X. The energy calculation uses a modular mean for trajectory direction and visual attention direction. Similarly we measure the information entropy of the distribution as:

$$H(X) = -\sum_i X(i) \log X(i) \tag{5.6}$$

Given the energy and entropy of distribution X we calculate the total similarity between any two Quad Tree nodes Q_1 and Q_2 as:

$$S(Q_1, Q_2) = \sum_f |X^f_{Q_1} - X^f_{Q_2}||E^f_{Q_1} - E^f_{Q_2}||H^f_{Q_1} - H^f_{Q_2}| \tag{5.7}$$

Where f is an index for each feature; speed, direction, and visual attention. We calculate the similarity between every two QT nodes in the scene giving an affinity matrix. We do not cluster the matrix at this point, instead the QT similarity becomes part of a distance measure in our behaviour metric outlined later. The method of comparing every QT node to every other suffers from poor scalability as it expands as $O(n^2)$. This can be reduced for large areas, or very dense QT structures, by limiting the comparison to a local region around a QT node only.

5.4.4 Visualising Scene Context

Having calculated the similarity between every QT node, and given the spatial clustering of similar motion it would be expected that neighbouring nodes have high similarity. The interesting observation is where neighbouring nodes do not have high similarity. Such cases reveal the scene structure as there are local spatial divides. In

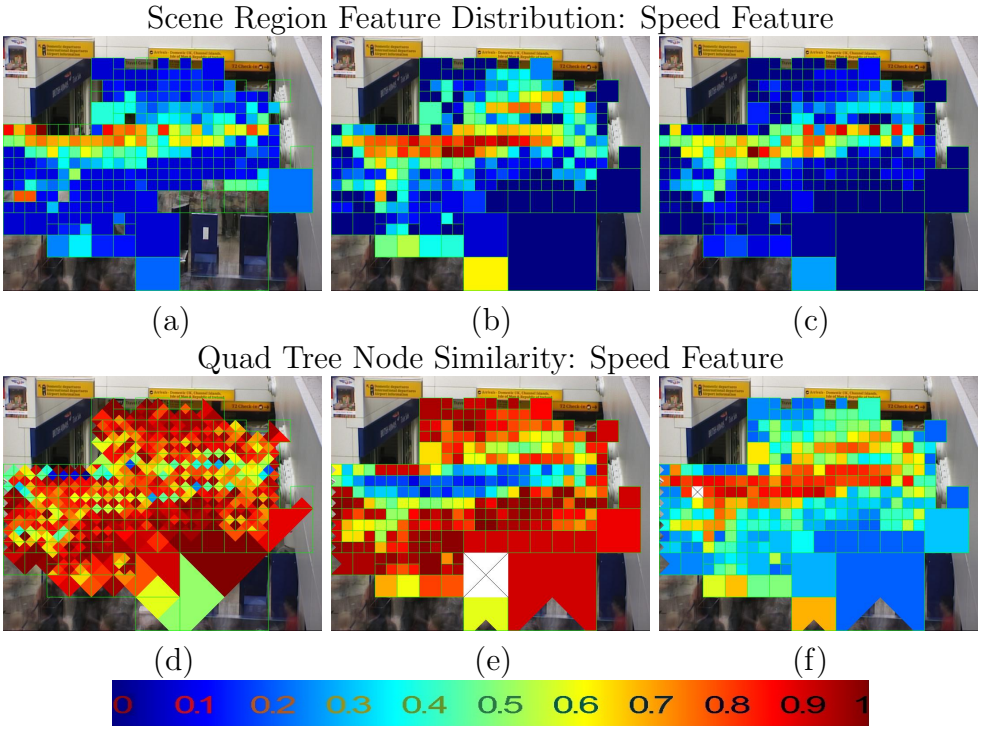


Figure 5.10: The colour bar (bottom) illustrates the mapping of colour to feature intensity (a) and energy (b) entropy (c) or similarity measure intensity (d). Image (a) illustrates the mean speed for the distribution of speeds in each quad tree node. The colour intensity map given in (e) ranges from cool to warm where warm indicates higher values in the normalised feature map. We observe a high amount of structure in the speed feature through the scene. A central horizontal band marks the appearance of a traffic lane through the scene. The neighbourhood similarity. image (d), illustrates regions of dissimilarity which indicate the underlying structure too the scene. Images (e) and (f) illustrate the similarity to quad tree node 224 image (e), 264 image (f). The quad tree node selected in each image is indicated in white with a black cross through the node. We see, unsurprisingly, that the quad tree node selected from the low speed region of the scene (e) has a low similarity to the central horizontal high speed region. However the quad tree node selected from the high speed middle band has low similarity to the rest of the scene, other than the central band, image (f).

the plot of QT node similarity based on speed we see such a structure emerge along the boundary of the high speed band in the middle of the scene, image (d). Given that a node can be compared to any other node in the scene within the behaviour space it is informative to see not only the similarity of each node to its neighbour but the similarity of a node too all other nodes in the scene. Below we illustrate for a selection of nodes their similarity to all other nodes. The selected node is indicated in white with a black cross. The similarity is likewise indicated by a colour intensity. Image (d) in the illustrations of scene features, Figure 5.10, 5.11, and 5.12, represents the neighbourhood similarity for QT nodes by displaying the similarity of each QT node to its neighbouring 4 nodes by the means of a triangular coloured segment. The left facing triangular segment of any node shows the similarity this node has to the node directly left of the node. Given the symmetry of the similarity score, neighbouring equal sized nodes take on a diamond like appearance.

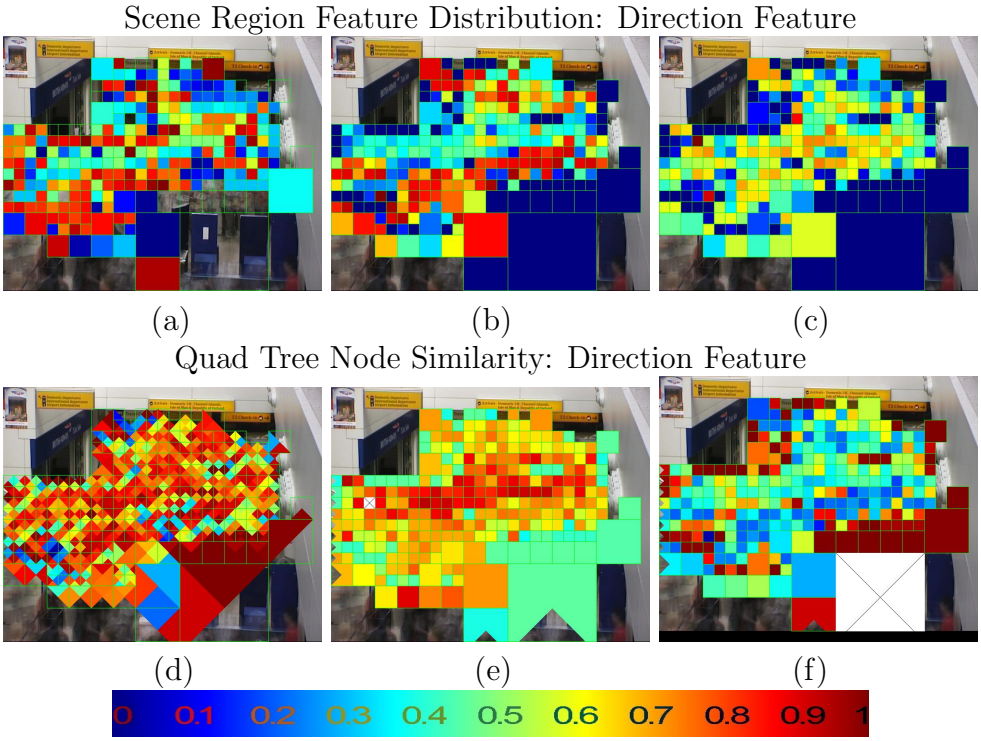


Figure 5.11: The colour bar (bottom) illustrates the mapping of colour to feature intensity (a) and energy (b) entropy (c) or similarity measure intensity (d). Image (a) illustrates the modular mean direction for the distribution of directions in each quad tree node. We observe a high amount of structure in the direction feature through the scene. A band through the middle is dissimilar to its surrounding space and has lower entropy. Conversely the energy in the scene is remarkably uniform indicating that marks the appearance of a traffic lane through the scene. The neighbourhood similarity, image (d), illustrates regions of dissimilarity which indicate the underlying structure too the scene. We illustrate the similarity between quad tree node 264 image (e), and 399 image (f). The quad tree node selected is indicated in white with a black cross through the node in each plot.

We next illustrate the distribution and neighbouring quad tree node similarity for the direction feature. The direction feature represents the direction in which people move. As was done with the speed feature, the direction feature is illustrated with a colour intensity map. The major difference being that the direction is modular and thus the distance between any two observations is a modular distance over 2π.

We observe a far higher entropy in direction than that of speed with clear spatial divides. Image (a) illustrates the modular mean direction taken within each quad tree node. It is apparent that the modular mean direction varies a great deal, with a central band similar to the speed feature demonstrating a distinct directional structure to the scene. Image (b) shows a high entropy for the upper and middle band of the scene, however a low entropy for the central horizontal band two thirds up the scene. This finding further confirms the positioning of a structured high speed route through the scene. There is a very low entropy and energy area of the scene in the lower right hand side which corresponds to a static queuing area. Image (c) indicates that the energy of the direction distributions are fairly uniform across the scene, more so than the entropy, with the exceptions of the lower right perimeter. Image (d) shows the neighbouring similarities between nodes. A less

prominent division between the central horizontal band and the rest of the scene is still present, here there is much similarity within the band and low similarity on the periphery, again indicating an underlying structure to the scene fitting with the findings of the speed feature. Again we illustrate cases of similarity to a single node below. Of particular interest is image (c) which the self similarity along the mid horizontal band, corresponding to a faster moving crossing across the scene. Image (d) shows the similarity between nodes in the lower right of the scene.

5.4.5 Visual Attention

Similarly to the use in the social model we extend the existing scene model to incorporate visual attention analogous to motion information. Rather than using a measure of visual interest between tracked individuals, equations 5.3 5.2, we use the observable feature; head pose. Head pose in this scope has direction, rate of change, entropy of direction, and energy of direction. The features for head pose are analogous to motion direction and as such are treated so. The justification of the use of visual attention and head pose direction is given in the introduction to this chapter 5.1, however we reiterate here for clarity. Motion features carry ambiguity in their behavioural implication due to external environmental influences upon the individuals motion and the limitation of expression motion alone can provide, the visual attention of an individual is free from the same environmental influences and may betray behavioural intents or interactions that the motion is incapable of displaying. Furthermore, whilst motion features give an indication of past to current interaction with the environment, the visual attention of an individual may indicate the future intention of the individual. To incorporate head pose into the scene context the energy and entropy of head pose forms part of the feature space characterising the scene. The behaviour metric takes into account the measure of head pose energy and entropy. For a more detailed illustration of the behaviour metric see 6.5.

The visual attention feature is comparatively highly structured across the scene. The colour intensity map is modular allowing similar directions across the $2\pi|0$ discontinuity to appear similar. We observe distinct regions of differing mean head pose direction in image (a). A higher feature entropy is seen centre of the scene than the edges, similarly with energy. This is not unexpected as across the lower portion of the scene the primary activity is waiting for a check in desk which has a high impact upon attention, drawing attention from the queue. Additionally there is less data, perhaps contributing to a less entropic distribution. Whereas near the middle and top of the scene visual attention is less constrained and well populated. Furthermore as head orientation and body orientation are not independent to each other, a high direction of motion entropy in this area would likely result in a higher visual attention entropy. The visual attention energy appears highest in the centre region of the scene and dissipates towers the periphery of the scene. This result likely represents an inclination towards focusing centrally in the scene when entering and exiting the scene. The quad tree node neighbouring similarity shows local structure in many places, not indicating a strong single structure to the scene but many smaller divisions in the scene. We next illustrate the quad tree node similarity to all other nodes.

We find an underlying structure to the scene indicated by the three feature distributions and their energy and entropy. A fourth feature of persistence, the measure of how long individuals remain in the scene, was also analysed. We found

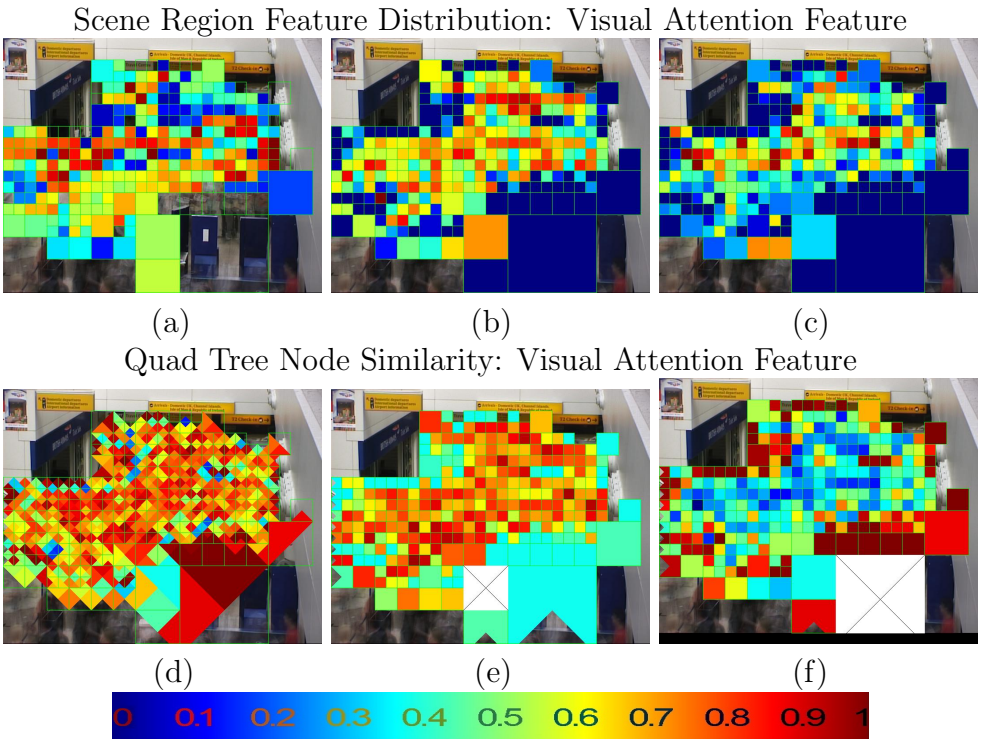


Figure 5.12: Figure (e) illustrates the mapping of colour to visual attention direction (a) and energy (b) entropy (c) or similarity measure intensity (d). Image (a) illustrates the modular mean visual attention direction for each quad tree node. We notice an interesting structure to the visual attention feature over the scene. There are regions of distinct conforming visual attention, such as top left and right on image (a). The lower right of the scene lacks in data samples, thus resulting in a low energy and entropy. We illustrate the similarity between quad tree node 224 image (e), and 399 image (f). The quad tree node selected is indicated in white with a black cross through the node in each plot.

very little distinction between regions in the persistence feature. For this reason we have not included the feature in our scene context equation and the results are not displayed.

5.4.6 Evaluation

Quantitative evaluation of the scene region location is ineffective as there is no defined ground truth. The feature exists only as a segmentation of behaviour space and thus to test the efficacy of the scene context we evaluate its merit by quantifying its impact upon behaviour analysis. We established previously in section 4.8 that scene context is a powerful tool in human anomaly detection. The use of the scene context with head pose direction in human behaviour anomaly detection is provided here 6.7. The above analysis of the features has however verified that there is a structure in the scene. This is most starkly illustrated by the speed feature, Figure 5.10, and additionally seen in the direction feature. The validation of scene structure indicates that the application of scene region context is appropriate in the behaviour metric defined in the next chapter. The structure located in the scene in question is intuitive to human observation when watching the video, adding further validation to the data driven approach taken.

5.5 Conclusion

In this work we built upon our previous work and recent advances in human behaviour surveillance to present an algorithm which models contextual information in a surveillance environment. Our approach is data driven and incorporates the feature of visual attention which is a particularly novel step using recent advances in coarse head pose estimation. Our approach successfully classifies social groups in the scene, achieving a true positive rate of 0.93 - 0.95 at a false positive rate of 0.02 - 0.05, depending on the dataset, and using ground truth visual attention cues. Using fully automatic feature extraction we achieved a true positive rate of 0.88 - 0.92 at a false positive rate of 0.04 - 0.07. The main contribution we make is the use of visual attention in a social estimation. We hypothesised that socially connected individuals display this through the visual attention feature by either looking towards each other or correlating attention. We have validated this by looking specifically for these two cases and improving upon a purely motion-based social clustering. This finding, and demonstration, opens a new methodology for automatic social estimation which may have implication beyond security; it may feature in marketing and crowd control analysis. We additionally developed our scene context beyond our earlier research, see Chapter 4 for our previous scene context framework. Our previous method used hard boundaries to classify the exact location of predefined regions. Our new method has no boundary definitions between regions and is not limited by predefined region definitions. Furthermore we include the feature of visual attention to enhance the feature strength of our system. Specifically, our contributions in this chapter are as follows:

- The use of automatic visual attention estimation in social group classification system for surveillance
- Evidence that social grouping is improved with the use of visual attention
- A method of deriving scene context information automatically, modelling the structure of the scene and comparative regional similarity

We next bring together all our previous research to develop a human behaviour anomaly detection system which uses the contextual information derived in this chapter, and visual attention, to detect behavioural anomalies.

Chapter 6

Detecting Abnormal Human Behaviour using Visual Attention

In this chapter we present our final human behaviour anomaly detection algorithm, NN-RCO, which builds upon all our previous research; incorporating in particular the social and scene context we previously developed. We detail the representation and features encoded for behaviour in section 6.2, how scene 6.3 and social 6.4 context information are used, and in section 6.5 we present the algorithm which detects abnormal human behaviour. In section 6.7 we demonstrate the feasibility and evaluate the proposed algorithm. We then provide a qualitative evaluation to the other state of the art techniques. This chapter addresses research objectives 5.2, 6.1 and 6.2; see section 2.7.

The work of this chapter and the data generated is published in Pattern Recognition Letters - Pattern Recognition and Crowd Analysis, 2013 [55], and in Computer Vision and Pattern Recognition conference 2014 [54].

6.1 Introduction

A problem facing the automatic detection of abnormal human behaviour is that of heterogeneous behaviour grouping. Particularly in unsupervised methods where observations of normal behaviour draw from multiple behavioural classes with no obvious segmentation and often a sparsity of examples preventing the learning of class boundaries in behaviour space; resulting in the masking of outlier behaviours. For this reason contextual information is be a powerful tool to provide class boundaries in behaviour space. The contextual information represents universally applicable a-priori expectations for the domain. Scene segmentation, social clustering, and temporal segmentation are three common examples of contextual information which enhance the interpretation of behaviour.

Previous research has shown that *context aware* anomaly detection based upon motion features alone is capable of detecting subtle abnormal behaviours, see Chapter 4. We now progress this line of research by bringing in the motion independent feature of visual attention. The intuition is that whilst motion features carry ambiguity in their behavioural expression due to external environmental influences, and are a limited form of expression, the visual attention of an individual is free from the same environmental influences and may betray behavioural intents or interactions that the motion is incapable of displaying. Furthermore, whilst motion features

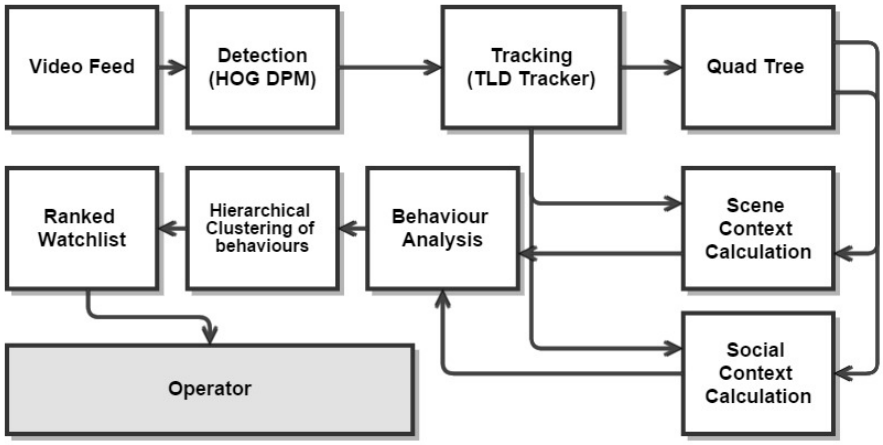


Figure 6.1: System diagram of our behaviour analysis technique. The process is mostly linear; processing datasets by batch. The tracking and detection precedes all other steps. Contextual information and quad tree calculation precede behaviour analysis. The operator sees only the output; the ranked watch-list.

give an indication of past to current interaction with the environment, the visual attention of an individual may indicate the future intention of the individual. Although we do not test this latter point explicitly; our focus is upon harnessing the visual attention profile information to better separate outlier abnormal behaviour from normal behaviour.

In this research we leverage our contextual information, the scene and social context, and apply visual attention to an alternative method to that implemented in our previous work, in Chapter 4. We seek to evaluate the effectiveness of each component of our system by its capability to detect abnormal human behaviour on several publicly available surveillance scenarios in which we manually groundtruthed the abnormal behaviour where not already available. We use the PETS 2007 dataset, and the Oxford dataset in our experiments as they offer complex dynamic scenes with subtle abnormal behaviours such as bag dropping and loitering. Before explaining and examining our method for context aware human behaviour anomaly detection we review relevant literature in anomaly detection using visual attention.

6.2 Behaviour Representation

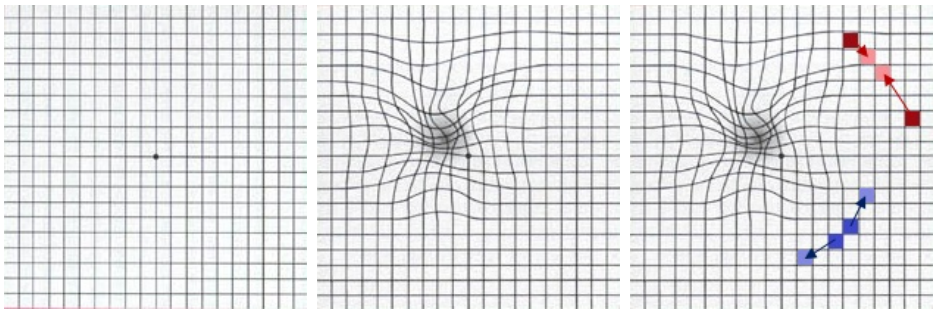
We base our methodology around the principle that non-trivial abnormal behaviours, those which cannot be detected by a simple aggregate motion dissimilarity, are difficult to detect because they are not clearly distinct from the tangle of normal behaviours populating the behaviour space. If we imagine all behaviours plotted on an arbitrary manifold, there will be a interspersion of normal and abnormal behaviours. Abnormal behaviours may reside on the perimeter of clusters of normal behaviour, or may even be intermingled with normal behaviour if the features describing the behaviours are not discriminative. In order to detect abnormal behaviours we would need to separate the clusters of normal behaviour from the relatively low frequency abnormal behaviours. In order to do this we need to carefully design a behaviour space surface which characterises behaviour in such a way as to create distance between abnormal behaviours, increasing intra-cluster distance, and reduces the inter-cluster distance, and thus amplifying the extent to which behaviour abnormalities

are outliers from the normal behaviour clusters. The key to our system's capability is in the selection of observable features from the data, and the inclusion of additive context features which will create intra-class distance between abnormal behaviour and normal behaviour. The feature space must be generic enough that it can be applied to general case human surveillance, and adaptive so that it can configure to the specific scene.

The start point for our algorithm is the observable feature space. This encodes any information that can be directly measured from the tracked image sequence. This consists of locations of pedestrians in each frame and an association between detections over multiple frames to encode a track. From here speed, direction, and head pose direction estimates can be drawn. Specifically for an image sequence $I = i_1, i_2, ..., i_t$ we draw out the set of detections $D = d_1^{i_1}, d_2^{i_1}, ..., d_n^{i_t}$ using the DPM [30]. Detections are associated between frames to form a set of estimated trajectories X where each trajectory $x_n \in X$ contains the speed v, direction Θ^D, position $p_{x,y}$, and head pose direction Θ^H at each time step t. We next include the contextual information and the quad tree coordinate system. Inclusion of the QT coordinates and contextual information can be seen as a change in basis to behaviour space. The behaviour space includes additional contextual information, further processed features such as visual attention estimates, and transformed features such as positional coordinates in quad tree coordinates. We first reiterate the aspects of the contextual information previously defined in Chapter 5.4.2.

6.3 Scene Context

We use a Quad Tree representation of the image plane to provide a meaningful coordinate system in which to represent position. When measuring the distance between any two observations later in our system we thus use a measure of distance based upon a coordinate system that is scaled by target density. The measure of distance also takes into account QT node similarity as defined earlier in section 5.4.2. This appeals to the notion of increasing interclass distance that we base our outlier detection behaviour analysis upon. We define later, that part of the measure of similarity between any two behavioural events is the proximity between the two events. However proximity is represented by two components; the quad tree distance scaled by node size, and the similarity of behaviour within the corresponding QT nodes the events come from. The former declares that distances in dense areas are exaggerated and distances in very sparse areas are contracted. The latter part of the metric ensures that behaviours have a low distance to other behaviours in similar regions, such as two queuing areas. This can be visualised as a warping of the image plane; expanding the image where there is a high density of tracks and contracting in regions of low density. Then, by including a measure of region similarity, we pull regions of high similarity towards each other and increase the distance between dissimilar regions. We illustrate this in Figure 6.2. Image (a) represents the Euclidean image plane coordinates the coordinates people are detected and tracked in. Image (b) represents warping of the coordinate system to accommodate information density. The coordinate system is transformed to present areas of high population in finer detail, where the exact coordinate system is defined by the quad tree representation. Image (c) represents the addition of region similarity. Those areas that are more similar are drawn closer together in the distance metric (Red blocks moving towards light red blocks) and those that are dissimilar have distance increased between them by the distance metric (Dark blue blocks moving away from



(a) (b) (c)

Figure 6.2: Image (a) represents the Euclidean image plane coordinates the coordinates people are detected and tracked in. Image (b) represents warping of the coordinate system to accommodate information density. The coordinate system is transformed to present areas of high population in finer detail, where the exact coordinate system is defined by the quad tree representation. Image (c) represents the addition of region similarity. Those areas that are more similar are drawn closer together in the distance metric (Red blocks moving towards light red blocks) and those that are dissimilar have distance increased between them by the distance metric (Dark blue blocks moving away from each other).

each other). This then represents our coordinate system where the distance between any two points is defined by the density of tracks and regional similarity score, as defined by our scene context work in section 5.4.2.

The scene context is in effect a soft boundary version of the previous defined scene context from Chapter 4 which split the scene into traffic lanes, idle regions, and convergence/divergence regions. Enforcing hard boundaries between regions causes discontinuities in the behaviour representation. Particularly when somebody traverses the order between two regions. In such cases the comparisons that can be made between the target in question and previously observed behaviours change rapidly resulting in discontinuities in the normality measure of the target. The soft boundary version of the scene context removes this problem entirely are presents a more reasoned approach to comparing behaviours from similar and dissimilar regions. Furthermore the soft boundary method does not require classification of regions which entailed previous definition for the regions to be constructed limiting the region classifications available. Instead the soft boundary method merely calculates the similarity of feature distributions within the regions spatially defined by the Quad Tree application. Thus new scene contexts can arise, which if distinct from existing regions by the nature of their feature distributions, will be automatically encoded into our system.

6.4 Social Context

We explain in depth our novel social context method in Chapter 5. Our method uses the appearance of motion dependency and visual interest between targets to cluster tracked people in a scene into social clusters. This method has been validated to give near optimal results against a social ground truth in the PETS 2007 data. We introduce the social model information in its full form with visual attention into anomaly detection at this stage of the algorithm; using it as a contextual feature to enhance anomaly detection. The social model is used in two separate but related

ways in our method. Firstly the social strength between two individuals increases the behavioural distance between those two distances. As such highly connected individuals have a high cost to comparing their behaviour too those they are connected to. The reasoning behind this restriction is to prevent a group of individuals all behaving abnormally from being self justifying. By introducing a high cost, which can be infinitely high, to comparing between socially connected individuals pedestrians must justify their behaviour in connection to those which are not motion dependent upon them. The second way in which the social model factors into the behaviour analysis is by propagating anomaly scores amongst a social group. In this way a the system has a notion of group abnormality. In its simplest form this can be merely applying the mean anomaly score to all members of a group. A less trivial application is to take a weighted mean where the weight depends on the strength of social connection, thus removing the need for hard classification of social groups altogether. Furthermore for some security applications it may be beneficial to extend a watch-list to individuals that have interacted with a target.

6.5 Defining the Behaviour Metric Space

The above social and scene contextual information, the head pose direction, and the motion features extracted and expressed in quad tree coordinates for the behaviour space. The behaviour space encapsulates our description of any behaviour observation. The dimensions in the behaviour space each represent a different factor that contributes towards the representation of behaviour. Our behaviour space is a 6 dimensional space, with orthogonal dimensions of: speed, direction, QT location, persistence, social group, and visual attention. Any observation of behaviour pertaining to a tracked individual at any given time can be described by a vector:

$$\mathbf{b} = \begin{bmatrix} v \\ \Theta^D \\ QT \\ \lambda \\ SG \\ \Theta^H \end{bmatrix} \tag{6.1}$$

which defines a point in behaviour space, where speed is denoted v, travelling direction as Θ^D, the Quad Tree location by QT, the persistence as λ, the social group identifier as SG and the head pose direction by Θ^H. We wish to extend the definition of behaviour space to a metric space such that any two points in behaviour space have a measurable distance between them. This is a critical step in the development of our system as it is determines the shape of the set of observations and serves as the basis for determining outlier behaviour profiles. The distance between point $\mathbf{b}_1$ and $\mathbf{b}_2$ is defined as:

$$\Delta\mathbf{b}_{1,2} = \frac{|v_1 - v_2|}{v_{max}} + \Delta\Theta^D_{1,2} + 1 - S(Q_1, Q_2) + \frac{|\lambda_1 - \lambda_2|}{\lambda_{max}} + \delta(SG_1 \neq SG_2) + \Delta\Theta^H_{1,2} \tag{6.2}$$

Where v_1 is the speed of observation 1 and $S(Q_1, Q_2)$ is the QT node similarity as defined in the Quad Tree node similarity equation 5.7. Persistence is a measure of how long the target Id has remained in the same Quad Tree location, as a normalised

value by dividing the value by the maximum persistence in the scene, and $\Delta\Theta_{1,2}$ is defined as:

$$\Delta\Theta_{1,2} = \frac{min(|\Theta_1 - \Theta_2|, |\Theta_2 - \Theta_1|)mod(2\pi)}{2\pi} \tag{6.3}$$

By this definition the distance between any two observations in behaviour space is the sum of the Euclidean distance between the speed of observation 1 and 2, the minimal modular distance between the direction of travel, the similarity of the quad tree nodes both observations here taken from, the difference in length of time each target has remained in the QT node, and the difference between the head pose direction of the two observations. Our metric space satisfies the four conditions of a non-negative metric space:

$$\begin{aligned} &(1) \quad d(\mathbf{b}_1, \mathbf{b}_2) \geq 0 \\ &(2) \quad d(\mathbf{b}_1, \mathbf{b}_2) = 0 : \qquad\qquad iff\ \mathbf{b}_1 = \mathbf{b}_2 \\ &(3) \quad d(\mathbf{b}_1, \mathbf{b}_2) = d(\mathbf{b}_2, \mathbf{b}_1) \\ &(4) \quad d(\mathbf{b}_1, \mathbf{b}_3) \leq d(\mathbf{b}_1, \mathbf{b}_2) + d(\mathbf{b}_2, \mathbf{b}_3) \end{aligned} \tag{6.4}$$

Thus satisfying the conditions for (1) non-negative distance between any two points, (2) preserving the identity of indiscernibles, (3) symmetry, and (4) triangular inequality.

The above description addresses the distance between any two points in the behaviour metric space. The trajectory of an individual x_n pertains to many observations in the metric space $\mathbf{b} \in x_n$ covering a distribution. This distribution forms the behaviour profile of a pedestrian, and the crux of our anomaly detection system. The distribution in behaviour space forms the total shape of the behaviour for any tracked individual, and the behaviour space metric provides a basis for determining the distance between any two behaviour profiles. However, in order to compare the behaviour profiles of two tracked individuals, we need to introduce a further metric to determine the distance between two distributions of observations.

6.6 Determining Behaviour Profile Similarities

For computational effectiveness we quantise the behaviour space into bins, reducing the behaviour space to a six dimensional histogram, which still satisfies the above conditions. Our previous work used a bin by bin similarity score between behaviour profiles to determine profile distance. This approach, similar to a correlation, was computationally fast however suffered from the draw back that it measured the degree of overlap between two profiles but not the total cross bin distance. In a case where to profiles have peaks in neighbouring bins there was a large distance score, and equally as large as when the peaks are very distant. Clearly this does not characterise the distance between two distributions particularly accurately. The solution to this is to use a cross bin distance measure. We select the Wasserstein metric, or Earth Mover's Distance, to measure profile distance as it provides a computationally efficient cross bin distance measure between two probability distributions. Intuitively the Earth Mover's Distance can be viewed as calculating the amount of work required to reshape a distribution into another distribution, here a unit of work corresponds to transporting a unit of 'earth' and unit of 'ground distance'. We wish to find the minimal work flow to redistribute behaviour space distribution $P = \{(p_1, w_{p_1})...(p_m, w_{p_m})\}$ to distribution $Q = \{(q_1, w_{q_1})...(q_m, w_{q_m})\}$ where in p_1

is an element in the behaviour space and w_{p_1} is the associated weight. The distance between any two clusters p_i and q_j is given by $\Delta\mathbf{b}_{1,2}$ which is calculated via the behaviour space metric defined in equation 6.2. Finding the minimal work flow $F = [f_{i,j}]$ between P and Q, where $f_{i,j}$ represents the work flow between p_i and q_j, can be solved as a linear programming task given $D = [\Delta\mathbf{b}_{i,j}]$ the distance matrix between any two points in behaviour space. We wish to minimise:

$$WORK(P, Q, F) = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j}\Delta\mathbf{b}_{i,j} \tag{6.5}$$

which is subject to the following constraints:

$$\begin{aligned}
&(1) \quad f_{i,j} \geq 0 && 1 \leq i \leq m, 1 \leq j \leq n \\
&(2) \quad \sum_{j=1}^{n} f_{i,j} \leq w_{p_i} && 1 \leq i \leq m \\
&(3) \quad \sum_{i=1}^{m} f_{i,j} \leq w_{q_j} && 1 \leq j \leq n \\
&(4) \quad \sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j} = min(\sum_{i=1}^{m} w_{p_i}, \sum_{j=1}^{n} w_{q_j})
\end{aligned} \tag{6.6}$$

Constraints (1) enforces a one way movement of work, which is to say you if you apply work to move a unit from P to Q you cannot then move a unit from Q to P. The second constraint limits the units that can be moved from p_i to q_j to at most the weight w_{p_i}, ensuring that the work done represents the total mass of the distribution and distance travelled. Constraint (3) preserves the symmetry of constraint (2). Constraint (4) ensures that the maximum units are moved. We use linear programming to find the minimal work flow F between P and Q via minimisation of 6.5. We can then calculate the Earth Mover Distance (EMD) as the normalised total flow:

$$EMD(P, Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j}\Delta\mathbf{b}_{i,j}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{i,j}} \tag{6.7}$$

which can be seen as the minimal set of distances between entries in behaviour space P and behaviour space Q. The EMD is normalised to ensure no bias towards smaller distributions. The EMD gives us a distance measure between any to distributions corresponding to the behaviour profiles of two tracked pedestrians.

As we mentioned previously, the key to our method is in a meaningful definition of the metric $\Delta\mathbf{b}_{i,j}$. The objective of the metric is to create distance between distributions which have different motion patterns, where witnessed across dissimilar areas of the scene, or are drawn from the same social group. The behaviour metric encodes the contextual information in this way; creating distance between distributions from different scene contexts, as determined by the Quad Tree similarity score 5.7, and preventing behaviour matching within a social group. This frame work is easily extended to additional sources of contextual information. The only requirement is knowledge of how similar one state in the context is to another. In the maritime implementation of this work we include the additional contextual information of ship class and tidal information. A similarity matrix crafted by expert knowledge is used to determine whether comparisons can be made by the different

classes of ship transmitting AIS signal, such as; tug, trawler, pleasure craft, and even land sea rescue helicopter. By this means we create insurmountable distance between distributions from classes that should not be compared, such as helicopter and trawler (and create similarity between those that can). Similarly we add a temporal context with the use of tidal context. We expect different distributions of behaviour to be witnessed at high tide and low tide as low tide imposes further restrictions upon the movement of ships.

The motion information is captured by three features, speed, direction, and persistence. Speed and direction intuitively capture the pattern of movement within the image plane, which is then referenced to a particular location by the quad tree coordinate. By this measure we characterise normal speeds and directions observed in any given location. The persistence feature plays an important role in time ordering the sequence of directions and speeds observed. This can be imagined as taking a 2D plot of motion vectors and stretching this into the third dimension to make a string of motion vectors, thus encoding the sequence of events, which is a powerful tool when comparing behaviours. By the nature of our behaviour metric there is an elasticity to the sequence of events; allowing similarity with a small reordering of events when making comparisons between behaviours. However bigger reordering of events comes at a greater cost. The optimal ordering of events being calculated by the EMD calculation as above. We treat the final non-context feature, visual attention, analogously to direction. In any Quad Tree location there is an expected sequence of visual attention direction, the persistence feature once again imposing ordering of the sequence. The power of the feature is its independence to motion and environmental impacts on motion, higher freedom from environmental constraints and thus less ambiguity.

Given the EMD method of calculating the distance between each behaviour profile we can then create a distance matrix $D = [d_{i \in I, j \in J}]$ which gives the distance between any two behavour profiles.

6.6.1 Creating a Ranked Watchlist

Given the distance matrix D we are in position to define outlier behaviour profiles. An outlier profile will represent the a tracked pedestrian which does not have a corresponding similar representative behaviour profile, where distance is defined in the above way taking into account visual attention, contextual information, and motion. A simplistic approach to defining outliers would be to take the profile with the greatest minimum distance to any other profile, or the greatest mean distance to the total set of behaviour profiles. For a trivial scene with a single type of behaviour these methods may be appropriate as the presuppose only one group of behaviours, the closeness to the centre of which are the most frequent most similar behaviours, the outliers being furtherest out. These naive methods fail to take into consideration heterogeneous behaviour spaces which consist of multiple behaviour types. In such cases there may be small clusters of legitimate behaviour which as a group are distant from the main cluster of frequent normal behaviour.

We turn towards Hierarchical Clustering as a means of determining the determining the outliers from the set of all behaviour profiles. We use the agglomerative 'bottom up' form of Hierarchical Clustering which presupposes every behaviour profile to be un-clustered and builds up clusters until all profiles are clustered into one group. The resultant 'greedy' clustering can be viewed as a dendrogram. The distances between profiles used in the clustering is given from the distance matrix D.

Hierarchical clustering requires a choice of linkage criteria to determine how the pairwise distance is measured. We use nearest neighbour clustering, also known as single linkage clustering. The distance between any two clusters is given as:

$$\Delta(A, B) = min_{a \in A}^{b \in B} D(A, B) \tag{6.8}$$

Where at each iteration of the clustering the two clusters A, B with the shortest distance are clustered based upon the minimum distance between any two elements a, b from the two clusters. Hierarchical clustering reveals the cophenetic distance at which any single behaviour profile is grouped with a cluster. The degree that any behaviour profile is considered an outlier is defined as the cophenetic height at which the singular behaviour profile was grouped into a behaviour cluster. Thus we can order the behaviour profiles by how great an outlier they are, creating a ranked watch list of behaviours, where each behaviour relates to a single tracked target. This is slightly different from anomaly detection as we do not classify which behaviours are and are not anomalies, but instead draw the attention of the operator to the most abnormal behaviours.

6.7 Experiment

To validate our approach we compare our systems ability against a ground truth list of anomalies for 3 different scenes. To illustrate how our system behaves we address the systems response to noise, and the impact of the addition of visual attention. A quantitative comparison to the state of the art systems would have little value. At the time of writing there is no comparable method which seeks to detect long term surveillance anomalies. Differences in representation such as the time scale at which behaviours are considered, whether behaviour pertains to the physical appearance, the short actions, interactions, or as in our case aggregate behaviour. Whilst we detect valid examples behaviours hat evolve in novel ways, there methods detect short instances of abnormal deviation, or patterns in crowd motion, or apply rule-based systems geared towards detection of a more nuanced behaviour than our statistical method. For this reason we provide an in-depth qualitative comparison to other methods in order to present our method in the context of other work in the field.

6.7.1 ROC Analysis

We first present the results of our system as a ROC curve to suggest the use of our system as a anomaly classification tool, where classification would be based upon whether or not the score is above a certain threshold. The ROC curve analysis illustrates the tradeoff between sensitivity and specificity for all possible thresholds. In real world applications a different pay-off of true positives to false positives may be desired. The ROC curve assesses the model independent of the choice of a threshold.

We first illustrate the true positive and false positive rate achieved when classifying the existence of abnormal behaviour in our 3 surveillance scenes. The scenes selected were PETS scene 4, PETS scene 0, and the Oxford dataset. The Oxford data is different to the PETS data in that it presents a far simpler behaviour set. There is only one type of behaviour, that of walking up and down the high street. The anomalies in the Oxford dataset consist of a man walking into the scene and standing stationary, and a man walking into the scene and using a bin. Note that

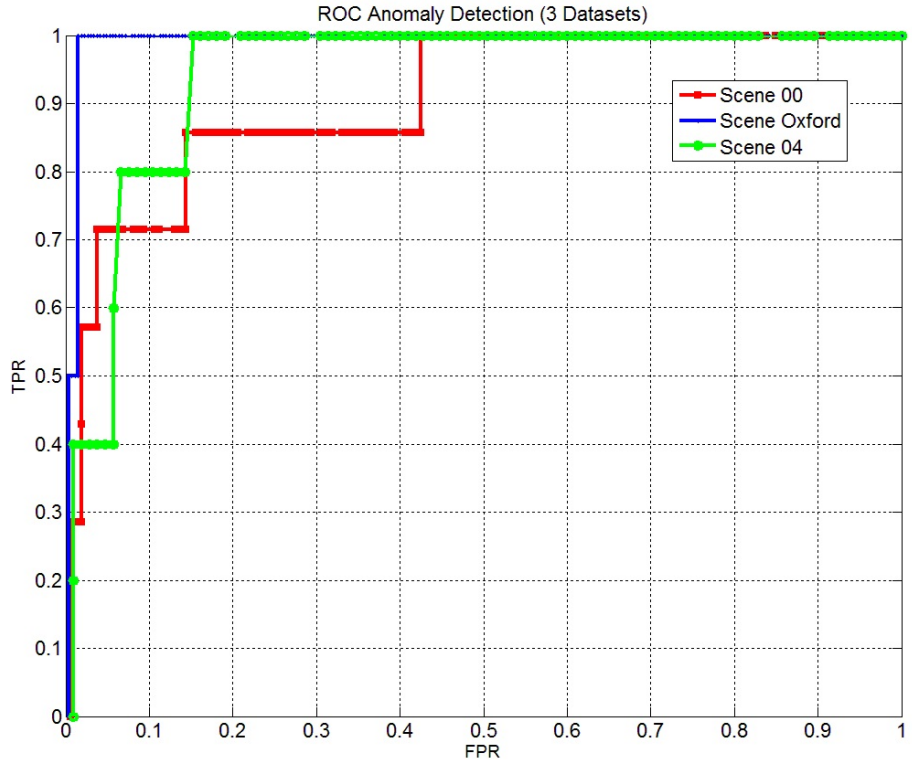


Figure 6.3: The TPR and FPR results for classifying anomalies in the PETS 2007 scene 00 data, scene 04 data, and Oxford dataset.

these behaviours are anomalies for the scene, although not a security concern. Our system selects anomalies based on their status as a statistical outlier; it does not make a threat assessment. Instead, a human operator in the loop would assess the event for the degree of threat it presents. The PETS scenes were selected for the complexity of the scenes. There is no typical behaviour of the scene. Dominant behaviours include moving swiftly through the centre of the scene, queuing, turning round a corner at the top of the scene. The anomalies in the scene are four people loitering; a motion which is well represented in other areas of the scene. Thus finding the anomalies is a hard task. We illustrate below the receiver operator characteristic of our system in all 3 scenes without the inclusion of noise in order to validate the methodology, we later re-introduce noise to demonstrate a real world capability.

Our system ranks all instances of behaviours in a watch list rather than making a hard decision as to whether an observed behaviour is normal or abnormal. Every entry on the watch-list has a corresponding anomaly confidence score. This score ultimately arises from them earth mover distance. We normalise the scores in the watch-list to give an anomaly confidence to each entry in the watch-list. To populate the ROC chart we then progress an anomaly threshold from 0 to 1 in increments of 0.001. It is possible for two entries to have the same earth mover distance and thus the same anomaly score. In such cases where two items have the same score and one is a false positive and one a true positive the ROC curve makes a uncharacteristic diagonal jump. We illustrate here in Figure 6.3 the main result of this chapter, the true positive and false positive receiver operator characteristic for our anomaly detection system.

The results displayed are for the three main scenes that we use in our experiments; PETS 2007 scene 4 and scene 0, and the Oxford dataset. We find that the

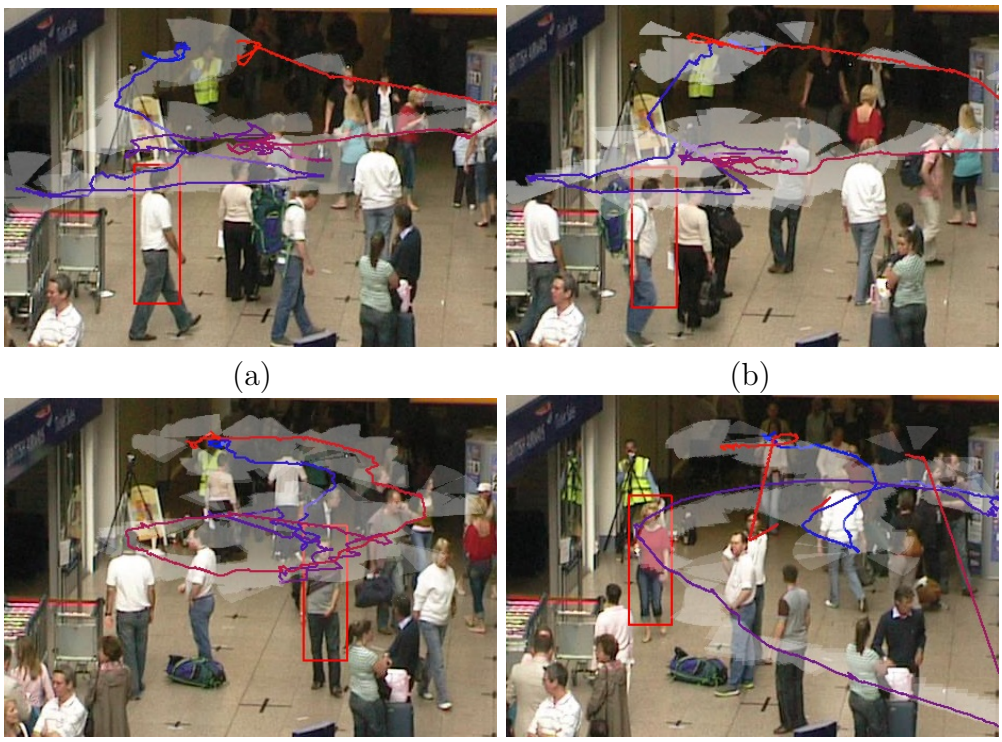


Figure 6.4: An example of two types of abnormal behaviour in the PETS scene 04 data. Trajectories are marked by a line tracking the head of the person, each image contains only one trajectory corresponding to one person. The colour indicate the sequence of the observations; blue earliest and red latest. Head pose is indicated by a pale field of view cone. In image (a) and (b) the two individuals in the red bounding boxes are engaged in a purposeful bag dropping scenario. In (a) and (b) the behaviour of loitering is being displayed. In images (c) and (d) the behaviour of loitering is being purposefully enacted. We are able to detect all four examples of this behaviour as abnormal in the scene due to their abnormal motion and visual attention patterns.

experiments in all three datasets have an initial true positive first response, ranking true anomalies as the most abnormal in the scene. In the Oxford scene the greatest anomaly corresponds to the unique behaviour of standing by and using a bin on the high-street. In the PETS scene 0 data the greatest anomaly corresponds to a young girl taking an abnormal route looking off at an unusual angle. The most striking result is the (blue) Oxford data which displays very few false positives. This data was selected for its scene simplicity; the anomalies are characterised by motion alone and normal motion is fairly homogeneous. The two motion anomalies (Interaction with scene objects, and loitering) are easily detected by our system. Both PETS experiments (red and green) have a greater number and greater complexity of behavioural anomalies than the Oxford data. Anomalies in both datasets range from loitering and other abnormal motion patterns to bag dropping and suspicious interaction. Both scenes display a similar efficacy achieving an optimal result of approximately 0.85 true positive rate at a 0.15 false positive rate. All three datasets detect all true positive abnormal behaviours before reaching a false positive rate of 0.45. The ROC suggests that an optimal payoff of true and false abnormal detections for all three datasets would be at the 0.4 true positive rate and approximately 0.01 false posi-

(a)

(b)

Figure 6.5: An example of an abnormal motion and visual attention pattern in the PETS scene 00 data. The tracked people in image (a) and (b) are socially connected, and display very similar trajectories. Both abruptly change direction in the middle of a normally one way direct route. Additionally the tracked child in image (b) spins around giving a very unique visual attention pattern. Trajectories are marked by a line tracking the head of the person. The colour indicate the sequence of the observations; blue earliest and red latest. Head pose is indicated by a pale field of view cone. The image on the left shows the trajectory in the image plane, the image on the right shows the same trajectory on the ground plane (birds eye view).

tives. We illustrate a few abnormal behaviours detectable by our system in Figures 6.4, 6.5, 6.7, and 6.7.

We perform an additional analysis on the Oxford data to validate the hypothesis that in a behaviourally homogeneous scene simply looking for the mean outlier distance of a behaviour to the single mass group of all behaviours allows us to accurately define outliers. To evaluate the hypothesis we calculate anomalies in the Oxford and PETS data via Hierarchical clustering and via simple mean difference. The results of this analysis are illustrated in Figure 6.8.

The results in 6.8 demonstrate that in the behaviourally homogeneous scene we test upon it is sufficient to identify anomalies by looking at each behaviour's mean difference to all other behaviours. This is equivalent to hierarchical clustering when there is only one cluster. However implementation of Hierarchical Clustering finds spurious false positive results which are deviations from the normal motion pattern, such as walking across the scene at a rare angle, however these anomalies are not globally the most abnormal. This demonstrates a danger of our system; in simple

(a)
l

Figure 6.6: An example of an abnormal motion in the PETS scene 04 data. The three tracked people run through the scene, one of which looks behind for much of the trajectory. Person 332 and 333 abruptly change direction in the middle of a normally one way route. Trajectories are marked by a line tracking the head of the person. The colour indicate the sequence of the observations; blue earliest and red latest. Head pose is indicated by a pale field of view cone.

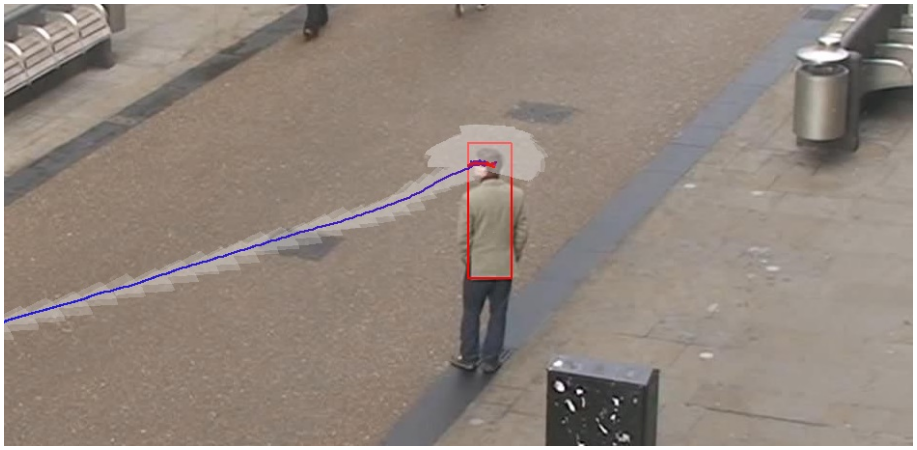


Figure 6.7: An example of loitering behaviour in the Oxford dataset. The individual carries out a rare example of stationary behaviour. Whilst not suspicious, the behaviour is abnormal in its uniqueness. The trajectory is marked by a line tracking the head of the person. The colour indicate the sequence of the observations; blue earliest and red latest. Head pose is indicated by a pale field of view cone.

scenes, using the more complex hierarchical clustering to detect anomalies forces too fine a segmentation of the dataset, emphasising subtle anomalies over the global more stark anomalies.

6.7.2 Impact of Feature Noise

We next illustrate the impact of noise upon the system, changing the ground truth social model and ground truth head pose direction for automatically generated features of social grouping and head pose direction, generated as stated in section 3.4

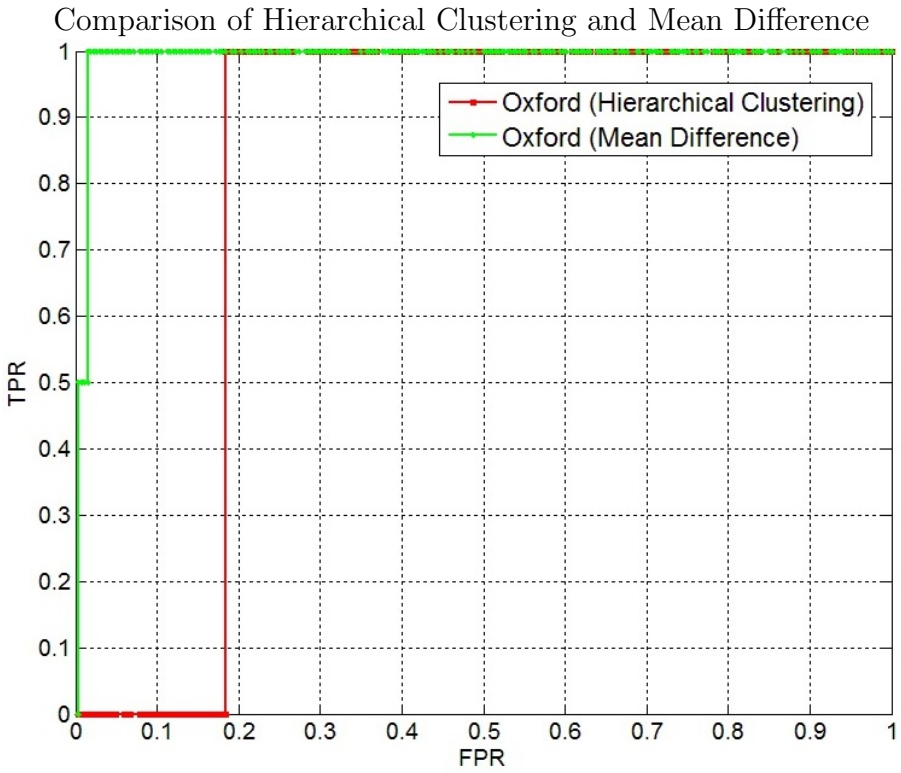
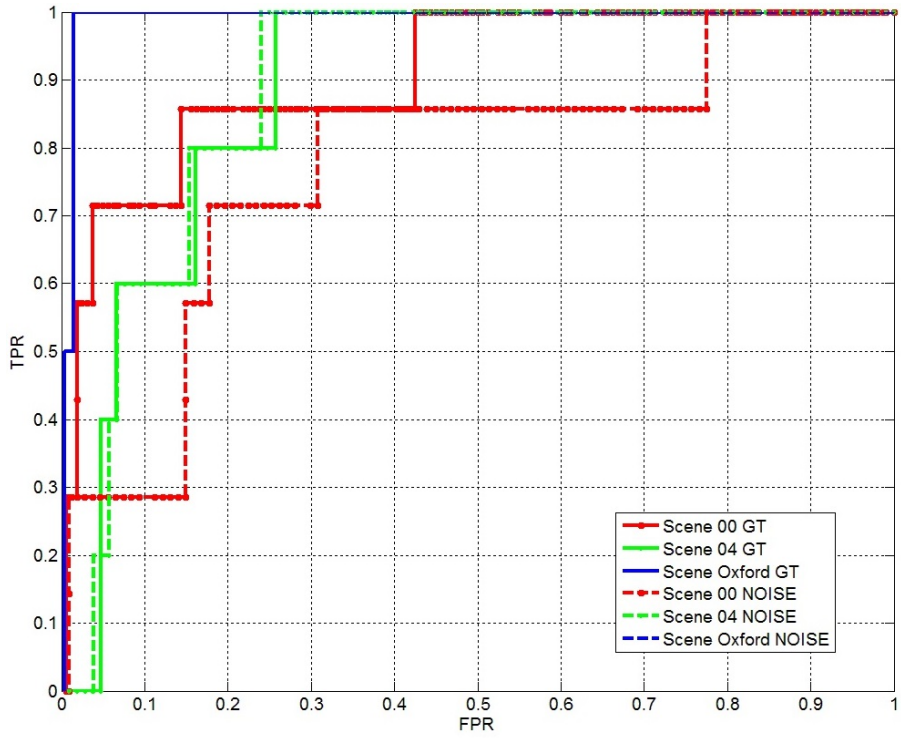


Figure 6.8: A comparison between using the Mean difference and Hierarchical clustering to determine outliers in a behaviourally homogeneous dataset, Oxford data [12].

and 5.1. This enables an understanding of the efficacy in a real world implementation. We illustrate the results in Figure 6.9. All three datasets return very different responses to noise in features. Motion tracking noise is unavoidable as is included in all experiments. We observe that there is no noticeable difference in TPR or FPR for the Oxford data. Both true positive anomalies are identified at equally as low a FPR as without noise. We put forward the idea that this is due to the anomalies being largely motion anomalies, and both anomalies not being in a social group. Thus noise in these two features has a low impact. We observe that noise in the PETS scene 4 data has a very slight impact increasing and decreasing the FPR by approximately 0.01 along the no-noise ROC curve. The scene 0 data shows the greatest response to feature noise, greatly increasing the FPR after the initial 0.3 true positives have been detected. Analysis of the false positives returned above 0.3 shows that their is suppression of several anomalies which should be linked in a social group however are not detected as the automatic social grouping fails to detect the group. Because anomaly magnitude is shared across social groups as detailed in section 5.2 the loss of social groups prevents association between these abnormal individuals lower scoring those in the group with lower anomaly scores. Furthermore, the inverse is true, those falsely high scoring anomalies in groups are suppressed when sharing scores across the rest of a normal group. Noise in the social grouping can break this suppression. We successfully demonstrated that social context has an impact on anomaly detection in Chapter 4 we see the impact of a noisy social grouping here, effectively muting the social context's efficacy. However, noisier real feature extraction does not incapacitate the system. We notice little or no impact on our system for two out of three of our datasets. This is perhaps due to the fact



Anomaly detection with noise

Figure 6.9: Illustration of anomaly detection efficacy, contrasting the results with noise and with minimal noise. For each scene we run anomaly detection with groundtruth features for social connections and visual attention, and with automatically extracted features. We see that only in Scene 00 is there a significant difference.

that our Behaviour Space representation of the features is malleable 6.5, allowing for skewing and time warping at a cost, thus making it possible to accommodate some noise. Furthermore, the Behaviour Space representation is a aggregate distribution which naturally hides high frequency noise, giving our system some intrinsic noise suppression.

6.7.3 Visual Attention Analysis

In our previous work we validated the use of contextual information in a motion-based anomaly detection system. We next evaluate the impact of visual attention upon the overall efficacy of our motion and context system. We achieve this by excluding the visual attention component and re-scoring the system against the ground truth. In this way the drop or gain in capability of the system can be measured. We demonstrate the impact of introducing visual attention upon the PETS scene 4 data. Note we use different settings than previous experiments in order to emphasise the impact of head pose. We witness a higher ranking of the first two ranked anomalies, and a large improvement in detection of the fourth and sixth highest ranked anomalies in the data. Analysis of the output shows the system ranks the anomalies in the same order, but with the inclusion of fewer false positives when visual attention is included. We demonstrate here 6.10 that visual attention has a beneficial role upon anomaly detection in this dataset. Thus we validate the intuition that abnormal behaviour is often partially characterised by an abnormal

visual attention in the scene. It must be noted that the head pose of a human is highly constrained by the body orientation, and has been shown to correlate with direction of motion quite highly [11]. For this reason it is possible that the visual attention feature is correlated with direction of motion and as such encodes much of the same information; possibly contributing to anomaly detection amplifying the motion in the analysis. Furthermore if the target demonstrates abnormal motion the visual attention must also be abnormal as it is required to follow the direction of the body.

We additionally we show on the simpler scene Oxford where the anomalies are starker in the motion component that head pose does not contribute particularly, see Figure 6.11. We find that the inclusion of visual attention reduces accuracy by higher ranking a small number of false positives. There is only a slight difference of approximately 0.01 higher FPR, corresponding to an additional 2 false positives. This however corresponds to an increase of 100%. The first anomaly to be detected, of the two in the scene, is the man using a bin for a prolonged period, this anomaly is not impacted by the inclusion of visual attention and scores equally as highly. The second anomaly however is weighted lower. The second anomaly consists of a man walking into the scene and stopping still. The head pose of the individual moves within a single quadrant slightly off from the direction of travel. This is altogether not an abnormal head pose, although the variance on the angle is moderately high for the scene. It is however very much an abnormality of motion not head pose that distinguishes this anomaly. Due to this we believe the inclusion of head pose masks the anomaly by weighting it lower due to the not-abnormal visual attention pattern. Although the anomaly is still very highly ranked in the dataset, this none the less represents a failure mode of the system. The inclusion of head pose information reduced the quality of the anomaly. The higher weighted false positives include a very short track of a women walking in a normal pattern in the corner of the screen, and an unusual trajectory entering the scene from an angle and avoiding an obstacle.

6.7.4 Qualitative Comparison to State of the Art

We have not found any work published to date which has attempted a similar task on data resembling our own looking at similar human behaviour as our own achievements. For this reason a quantitative comparison to state of the art methods is unattainable. However a study of similar methods has led to a qualitative comparison of state of the art work allowing us to better locate how our work fits with the leading edge in the field.

The main barrier to comparing methods quantitatively is that difference in the abnormal behaviours sought (bag dropping, following, loitering, interaction between people, cars, human, maritime etc) and the nature of an anomaly (short term, long term, instantaneous) result in very few methods returning comparable anomalies. There is no benchmark test for anomaly detection as the definition is broad and encompassing of many different tasks.

We compare first to work by Xing [39] which seeks to find anomalies in semi crowded human surveillance using spatial and temporal context. This work uses a novel local nearest neighbour distance (LNND) descriptor for anomaly detection in crowded scenes. In brief, the method works by taking an input video and segmenting it into equally sized spatio-temporal cuboids. Each cube is an event in the video. The similarity between an event and its local surroundings is calculated by the Earth Mover Distance similarity measure of the distribution calculated by the

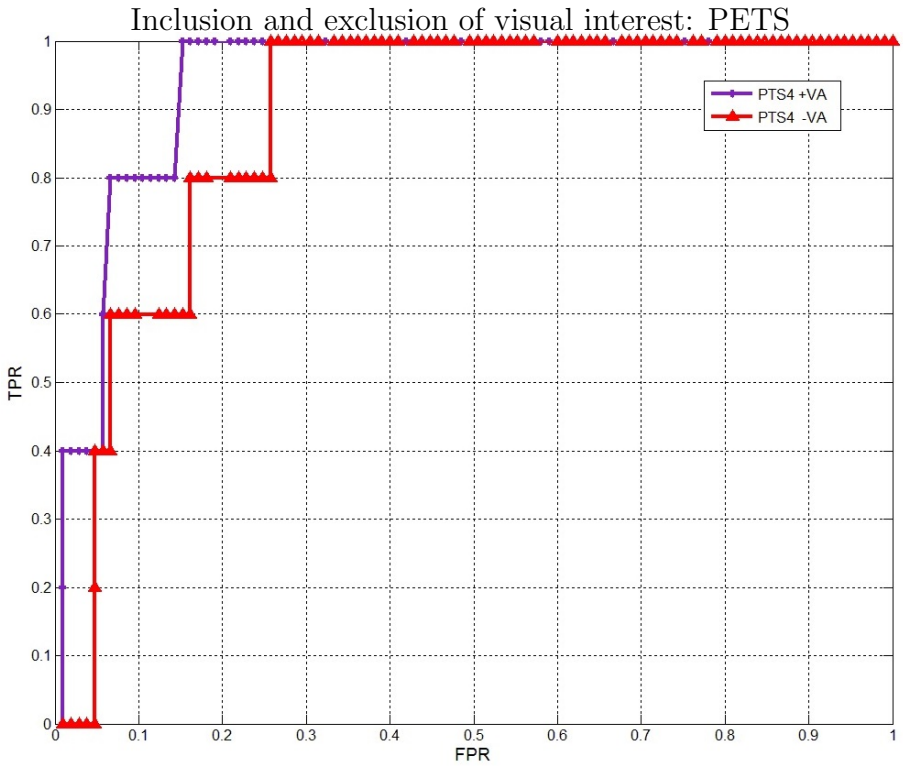


Figure 6.10: Illustration of the anomaly detection efficacy of our system with and without the inclusion of visual attention on PETS Scene 4. We find that visual attention permits the discovery of 4 out of 5 anomalies earlier than without visual attention.

Local Motion Pattern feature descriptor. Each event then has a similarity to its local neighbourhood events. Outlier events are considered to contain anomalous events. This work bears the similarity to ours that it attempts to address behaviour in the light of context. Events, which are the atomic element of behaviour in this work, are compared to those events that are spatially close by and temporarily close. In this way the work mirrors ours in that similarity can be found only in the confines of the local behavioural environment. In our work this is achieved by looking at the similarity of the location to other regions in the scene and within the same social class. Xing's work implies a strong spatial and temporal neighbourhood context, however the work encodes behaviour very differently to our method. The atomic element of behaviour in Xing's work is the spatio-temporal event whereas the atomic component of our behaviour analysis is a person with motion and visual attention for a frame. Our method is human centric; behaviour is at its most basic a human agent event. However, in Xing's work, events can pertain to multiple individuals or non-human events. The benefit of this approach is that you do not have to explicitly search for human targets; a costly stage in many systems. However, it weakens the behaviour representation and the descriptive capability of the system. Higher level processing often requires an understanding of the domain and the agents in the environment. An example of such is our social context work or the rule-based behaviour analysis of Robertson [74]. Ultimately if the behaviour representation does not encode an understanding of agents the analysis will be limited to motion-based understanding. Such an approach is very relevant for crowd for analysis however subtle behaviours such as loitering in a crowd or suspicious following will not be

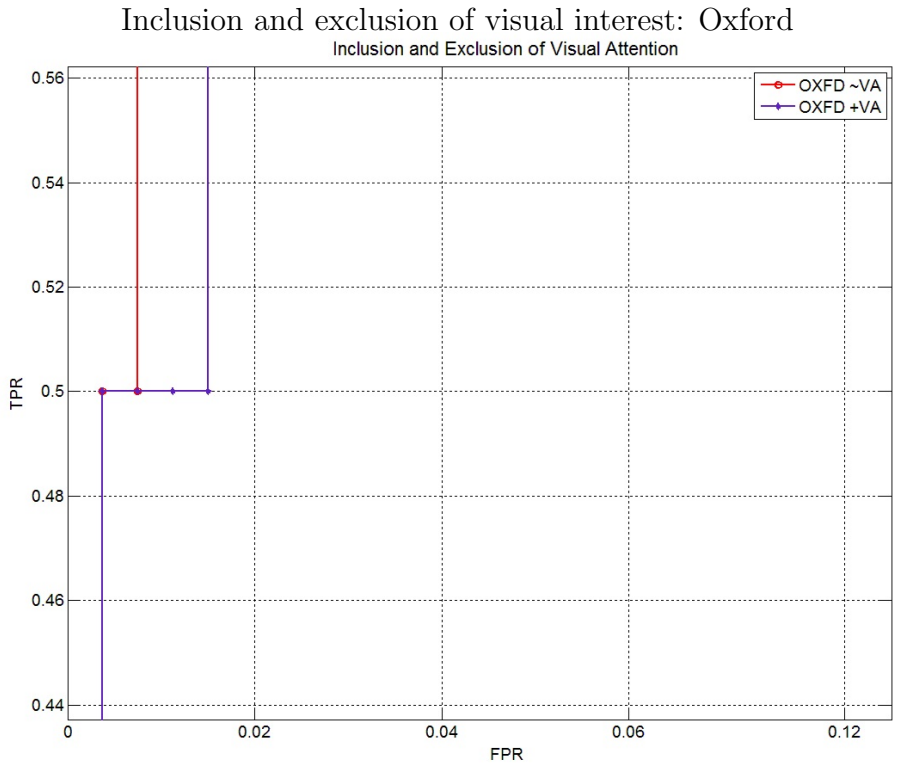


Figure 6.11: Illustration of the anomaly detection efficacy of our system with and without the inclusion of visual attention on the Oxford data. Note the scaling of the axis for better visualisation. We find that visual attention permits the discovery of the second anomaly in the scene earlier than without visual attention.

detected. This is observed when we analyse the anomalies detected by this system. The anomalies detected by this system include bicycles, small trucks, skateboarders, and a wheelchair user. The underlying feature in an event is the local motion pattern which in part encodes the appearance of the moving targets. Thus it is not surprising that an image patch containing an object of starkly different appearance and speed to the majority of individuals in the scene would be flagged as an anomaly. Such a method may be very appropriate for detecting crowd motion anomalies, however to detect more subtle agent behaviour such as loitering, following, or bag dropping this method would not be applicable.

We turn to the work of Loy [61] as a contemporary work comparable to our own. Loy aims to address the problem of anomaly detection of complex subtle and difficult to detect behaviours. In particular Loy addresses those behaviours difficult to detect due to the complex temporal dynamics and correlations among multiple objects behaviours. Complex behaviour is modelled using a cascade of Dynamic Bayesian Network (CasDBN) which capture its temporal characteristics or spatial-temporal visual contexts. Loy claims the cascade structure maps naturally to the structure of complex behaviour, allowing for more effective detection of subtle anomalies. In this work a video is manually segmented uniformly into non-overlapping video clips. Foreground is extracted and represented as a 10 dimensional feature vector, consisting of object centroid, width, height, occupancy, ratio of dimension, and mean optical flow. The system thus, at its most basic, represents motion only; characterising the position, shape, and motion of foreground activity in a video sequence. The motion of foreground blobs are considered atomic events which are

clustered by k-means into sets of atomic events corresponding to behaviours in the scene. The first stage of the CasDBN is composed of multiple Multi-Observation Hidden Markov Model (MOHMM)s, each of which is used to model the temporal sequence of atomic events within a single region. The MOHMMs are structured as a temporal hierarchy to capture behaviour that unfolds at different scales. The output of the first section is filtered and passed to the second stage which consists of a MOHMM for modelling the state sequences inferred from the first stage. Whilst the first stage models how behaviours typically unfold the second stage is responsible for learning the global correlations among local behaviours across regions.

This work is similar to our own; the behaviour representation considers the long term sequencing of events. The system is applied to a similarly large surveillance area, however applied to vehicles. Additionally the method is applied to sparse human surveillance in an indoor environment. Critically, however, Loy uses a contextual perspective to address the detection of subtle and complex abnormal behaviour. Events are seen in a local context in the first Dynamic Bayesian Network (DBN) stage and then assessed in a global context in the second. This approach enriches the behaviour representation, permitting the an analysis that goes beyond straight forward intrinsic motion anomaly detection. Whilst the work is fundamentally attempting to solve a very similar challenge it is dissimilar to our own work in functionality. Behaviour is a global motion pattern of 'foreground blobs' rather than pertaining to the individual agent. The approach explicitly models the transition probabilities between events, allowing for novel sequences, rather than modelling behaviour as a particular sequence. Importantly the method must be calibrated to the expected behaviours; the segmentation of the training video into events is done manually, the hierarchy of DBNs must correspond meaningfully to the temporal progression of behaviours to correctly encode transitions between actions. We overcome the temporal scaling of behaviour by implementing a flat approach which does not require scaling with time. We populate a distribution in a behaviour space representation and use the shape and position of the distribution as the defining characteristic of the behaviour. The use of a cross bin similarity score with a linear monotonic increase in cross bin cost provides a malleability to comparison between behaviour; allowing for some warping of the behaviour distribution. The benefit of Loys approach is primarily found in the robustness to occlusion and sensor noise due circumventing object tracking, and the reduced computational burden this approach incurs; the complexity of the DBN structure is only $O(Q^D T)$ compared to $O(Q^D T^3)$ of original Hierarchical HMM implementation. Where Q is the number of states in each layer, and T is the length of a sequence. Events are observed in the context of motion elsewhere in the scene enabling relations between regions to be encoded. However, over fitting may result in an overly restrictive model of the scene as unique but legitimate patterns across the scene could be low scored and dynamic changes in the scene are not adapted to. Furthermore the approach places a strong dependency upon a structured dependency across all regions in the scene, an assumption which may not be satisfied in more irregular environments than the traffic scene tested upon. Quantitative comparison to this method is infeasible as we cannot demonstrate the efficacy of visual attention or social modelling in the vehicle traffic scenes tested in this research.

6.8 Conclusion

In this chapter we bring together our previous work in context aware behaviour analysis, visual attention extraction and exploitation, and experience in maritime behaviour analysis. We developed and tested the NN-RCO system which successfully detects abnormal behaviour based upon the motion and visual attention that a target displays within the context of the different scene components and social surroundings. Our method returns alarms for a number of purposeful abnormal behaviours in the PETS datasets, and natural behavioural anomalies in the Oxford and PETS data. Of particular interest to security is the detection of loitering in a scene where waiting in some areas is commonplace, a behaviour characterised by motion and visual attention. Abrupt changes in motion, and novel trajectories through the scene are flagged to the operator, as well as unusual interaction with scene objects and bag dropping. We demonstrated the efficacy of our system by ROC analysis. We additionally analyse the impact of different components of the system, the importance of visual attention, and the impact of noise in the features. The main contributions made in this chapter are:

- The use of visual attention in a full human behaviour anomaly system
- A novel anomaly detection system capable of including context information and simply integrating additional features such as visual attention.
- A novel method for long term profiling of behaviour that elegantly handles tracking noise
- Evidence that subtle behaviours such as loitering and bag dropping have a visual attention element in their composition

We next test how well our algorithm generalises by applying the contextual nodes (social and scene context) and our NN-RCO algorithm to the maritime domain. By doing so we test to ensure the algorithm is not environment or domain specific in it's application but instead has implication in the wider scope of surveillance.

Chapter 7

Maritime Behaviour Analysis

We apply the lessons learnt from the context aware behaviour analysis previous carried out in Chapter 4 to the maritime domain. The purpose of this chapter is to demonstrate the versatility and application of our NN-RCO algorithm in an alternative domain. This chapter serves to test the generic applicability of our approach. We wish to avoid developing an algorithm that is over-fitted to our particular data, or works only by exploiting some nuance of the human surveillance data. We wish our approach to be able to handle a broad spectrum of different behaviours and behavioural variation. We address this, in part, by using an adaptive approach; creating a system that defines normality relative to what it has seen before rather than using templates. However to test that our system is generically applicable we need to test on data that is characteristically different to the human surveillance data we targeted our system at. We test upon the maritime domain in order to assess this. The algorithm used is the same as human behaviour analysis algorithm from Chapter 6 with adaptations outlined in section 7.3. The work we outline in this chapter derived from a real world application of our research and as such also demonstrates the impact of our research. Our primary source of data is the publicly broadcast AIS signal which presents GPS locations, speed, direction, and meta data for every ship within range. We capture the data with an aerial in house which gives range over Southampton and Portsmouth, in Britain. Our objectives for the maritime domain are the identification of suspicious behaviour in and around the background of legitimate traffic, apply algorithms capable of reducing operator workload, and to employ an algorithm capable of improving maritime situational awareness.

7.1 The Maritime Domain

Maritime behaviour, while agent-based, is different from human behaviour in its pattern of motion and time span. In particular the scale of the area to monitor is larger (10s of kilometres) making anomalies difficult to be seen, the scale of large movements cannot be appreciated simply, the motion of vessels is slow leading to an impression of lack of continuity in behaviour, often the motion is below the observation limit, and patterns are difficult to see when the repeat period is long. Abnormal and normal behaviour within the maritime domain is characteristically different to that of human behaviour. Abnormal behaviour may include events such as unexpected stops, deviations from standard routes, speeding, traffic direction violations, or novel motion patterns. Threats may include smuggling, sea drunkenness, collisions, grounding, terrorism, hijacking, or piracy [53]. In the maritime domain any ship over 300 gross tonnage is required to transmit an AIS signal at

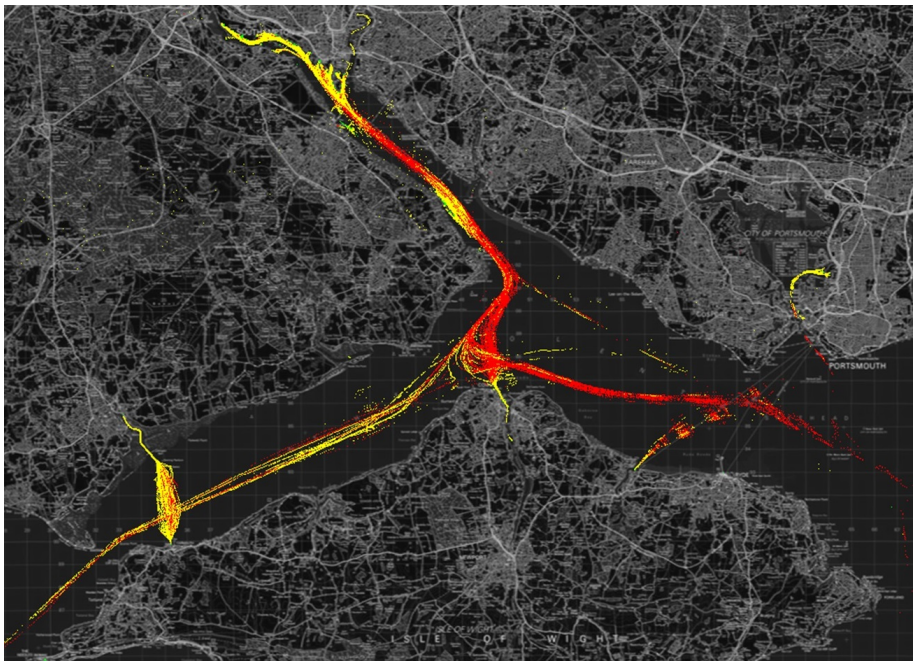


Figure 7.1: Illustration of all the data points collected in a 66 hour AIS data collection. The data points are coloured yellow to red over time per track. The plot reveals that ships tend to approach the port at the top centre of the image travelling from the left, and leave travelling to the right. The plot further shows blind spots on the lower right of the image, indicated by gaps in the tracks.

a periodicity relative to the velocity of the vessel. The signal is public, permitting anyone with the appropriate equipment to receive and log the data. We record our data with an antenna on top of the lab. AIS transceivers automatically broadcast information, such as their position, speed, and navigational status, at regular intervals via a Very High Frequency (VHF) transmitter built into the transceiver. The information originates from the ship's Global navigation satellite system (GNSS) receiver and gyrocompass. Other information, such as the vessel name and VHF call sign is transmitted at different regular intervals. The AIS signal encodes the unique numerical ID of the vessel enabling accurate tracking over extended period of time. It is these tracks that we use to detect abnormal maritime behaviour.

7.2 Background

We base much of our review of background literature upon the work of Laxhammar [53] and his findings. As a generalisation, within the maritime domain the algorithms proposed are accompanied by less information regarding implementation and the experiments are explained in less detail. Focus tends to instead be upon application and theory. The result of which is it is harder to reimplement and validate experimental results within the maritime domain.

We start by detailing methods which norm-based, or data-driven, methods. In Ristic et al. [73] they extract motion patterns from AIS data which are then used to construct motion anomaly detectors using kernel density estimation. Data from new trajectories is sequentially classified as normal or anomalous based on their likelihood, which is calculated from the Probability Density Function (PDF) of the

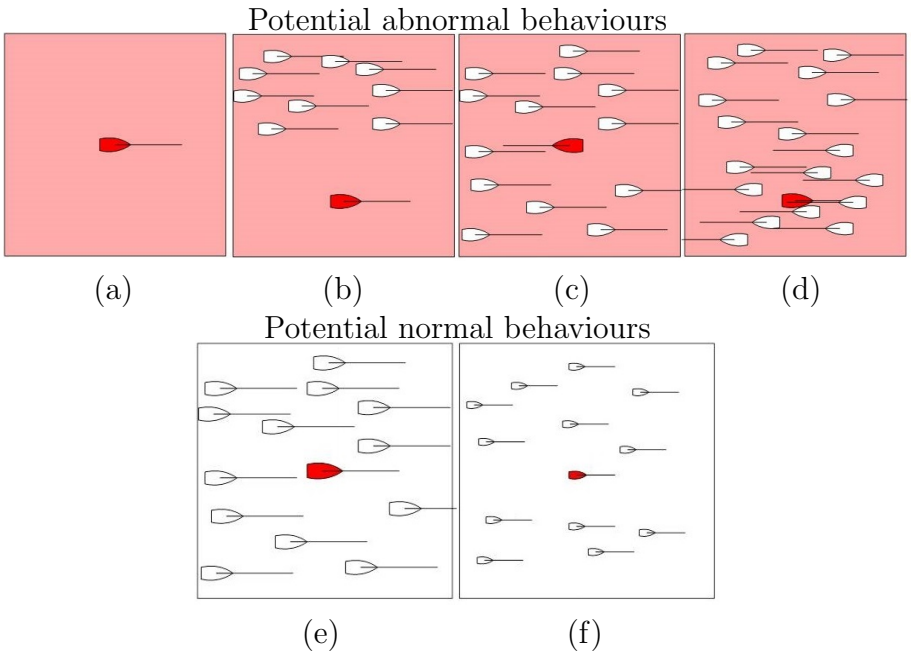


Figure 7.2: Illustration of 6 potential normal (white background) and abnormal (pink background) shipping behaviours. Images (a) through to (d) show abnormal behaviours. Images (e) through to (f) show normal cases. (a) shows a solitary ship in a region with no training data which may or may not be considered abnormal. it is however considered novel. Image (b) shows a ship travelling in a novel location with normal motion. (c) show a ship with contradictory motion (direction) in a normal location. Image (d) shows a harder case of contradictory motion where the motion is only abnormal to the positional context.

corresponding motion pattern. Prediction of trajectories can be carried out using the historic motion pattern data to build a Gaussian sum tracking filter. In Kraiman et al. [50] they propose and develop an Automated Anomaly Detection Processor that exploits data fusion (multi-INT), multi-sensor tracking and surveillance data to detect abnormal events in an unsupervised manner. The method uses a self-organising map to cluster input data and a Gaussian Mixture Model to model the clusters. The models are then used to classify anomalies based upon the probability output of the Bayesian probability output. One of the advantages of Bayesian reasoning in a system is that it enables the incorporation of domain expert knowledge [41]. Furthermore it provides the advantage of the possibility for humans to understand and interpret the learned model. Johansson et al. demonstrate their method upon synthetic data showing detection of simple cases such as speeding. Trajectory learning can be enacted using HMM and Gaussian Mixture Model (GMM) approaches. Urban et al. [84] use GMM to model the position data of trajectories based upon Expectation-Maximization (EM). Urban then uses the GMM models as states in a HMM which is estimated from the trajectory data using the Baum-Welch method. The likelihood of the trajectory can be drawn from the HMM and a threshold is used to determine whether the trajectory is an anomaly or not. The transition probability between regions is modelled by Tun et al. [83] in which regions are found using density maps and Linear scale space. A HMM is used to model and determine the probability of trajectories. Neural networks have often been used to model trajectory and contextual information in a data driven approach. Rhodes

uses supervised and unsupervised learning of a neural network in order to classify anomalies [72]. Rhodes et al. use a fuzzy ARTMAP neural network which allows for human input labels on clusters. Using the location and speed of ships new instances are classified as normal or abnormal depending on the nearest cluster. Bomberger et al. present an alternative use of neural networks that predict the future location of ships based upon the current location in a grid, speed, and heading. Abnormal behaviour is defined as a trajectory that deviates from the predicted route [15].

7.3 Application to Maritime

We apply our technique as describe in Chapter 6 to the maritime domain. The motion features, Quadtree representation, and social grouping all have direct application or analogous use in the maritime domain. However there is no equivalent for visual attention, and thus this feature is removed from the algorithm when applied to shipping data.

A difficulty for the detection of anomalies in maritime shipping is the wide areas that require monitoring coupled with large variations in shipping densities. For example the shipping lanes approaching ports are narrow with high traffic flows whilst the open ocean is largely empty. To represent position at high enough detail would be computationally intractable and ineffective in sparse open sea areas. Conversely to represent the position coarsely would not capture the detail required to spot anomalies in shipping lanes and ports. The use of the QuadTree representation, from section 5.4.1, of the ground plane addresses this difficulty in the maritime domain. The coordinate system thus represents the density of ships in a region and allows us to model sparse areas efficiently whilst retaining high positional resolution in dense areas.

The social model from section 5.2 has an analogous use in the maritime domain. We model motion dependency that arises from convoy behaviour or direct dependency such as tugs pulling ships. Additionally we witness similar motion from two ferries making short repeated trips, we class these as motion dependent. The dependency largely arises from the shared trajectory and the synchronization of arrival and departure times. The calculation for shipping motion dependency is identical to the human social grouping, with the removal of visual attention.

7.4 Experiment

We next illustrate the findings from applying the scene context and behaviour analysis to the maritime domain. We compare performance of the anomaly detector using a hand labelled ground truth data set. Social and scene context are not quantitatively evaluated, however the results are detailed. The dataset we use was recorded over 66 hours during a weekend using an AIS antennae on top of our lab. We recorded 199 ships over the period after excluding AIS signals from non-ship entities such as buoys, light houses, and land sea rescue helicopters. The average length of a trajectory over the time span was 17.6 hours, and the average speed was 7.52 knots, with a standard deviation of 0.45 knots. The data was collected over an area of approximately 55km wide and 40 high, with sporadic signals received beyond this. We confined signals to the $2200km^2$ area around Southampton, excluding any beyond this.

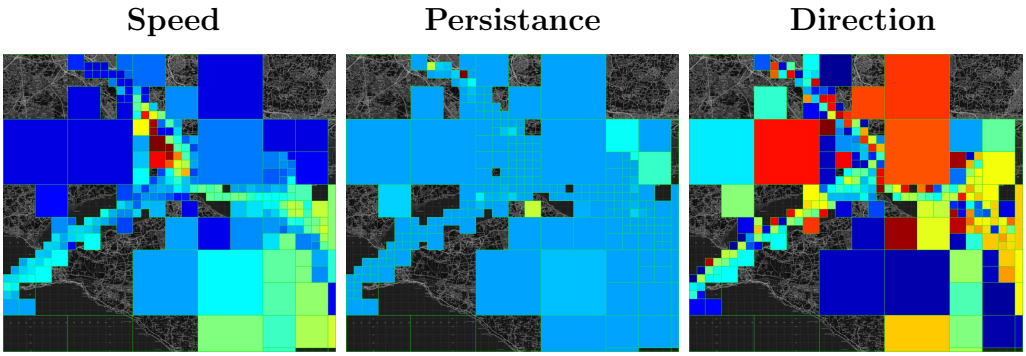


Figure 7.3: The mean feature intensity for speed (a), persistence (b), and direction (c) in the maritime 66 hour data set.

7.4.1 Scene Context

We can apply the scene context work from Chapter 5 directly to the maritime domain. Maritime ground plane coordinates are trivial to work with as all positional information is provided as GPS coordinates. We illustrate a subset of the observable features and derived features that compose the scene context for the sake of brevity. We seek to provide enough that the structure of the maritime scene can be established and the generalisation of our algorithm can be validated.

As previously stated, the use of the Quad Tree representation takes on a further use in the maritime domain. Within a port it is important to model the scene to a fairly high resolution in order to capture shipping lanes, docked areas, and convergence regions. However, in open ocean the same high resolution modelling would not be suitable, and would be computationally intractable. The QT representation gives us a way to model high information areas to a high resolution and low information areas to a low resolution. The QT is applied similarly to its implementation in human surveillance. QT nodes are divided into four sub-nodes when 10 unique ships are recorded in the node. The results of the scene context are as follows. In Figure 7.3 we illustrate three of the observable features that compose the scene context (Speed, persistence, direction). We observe that the mean speed over the scene is the most highly structured of the three features. The Solent water is the strait that separates the Isle of Wight from the mainland of England shown horizontally as a downward facing convex curve. It is about 20 miles in length and about one to four miles wide and a major shipping route for passenger, freight and military vessels.. We observe uniform speed across this stretch of water. The highest mean speeds are observed travelling up the Southampton water, vertically and centred, as this is a regular path for the Red Jet high speed ferry. The persistence in the scene is very uniform with the exceptions of the ports that show a far higher mean persistence. Direction is more chaotic with structure appearing only in the shipping lanes across the Solent and up Southampton water. To better appreciate the shape of the feature distributions attributed to these regions we show the entropy of each distribution in Figure 7.4. The entropy of the distribution for a given QT node provides information on how structured the motion of ships through that region are. Which, in turn, indicates how constrained behaviour is for that region. We observe that the speed feature is highly constrained along the convex horizontal and far less so travelling up the vertical Southampton Water. Thus we find that speeds are typically higher

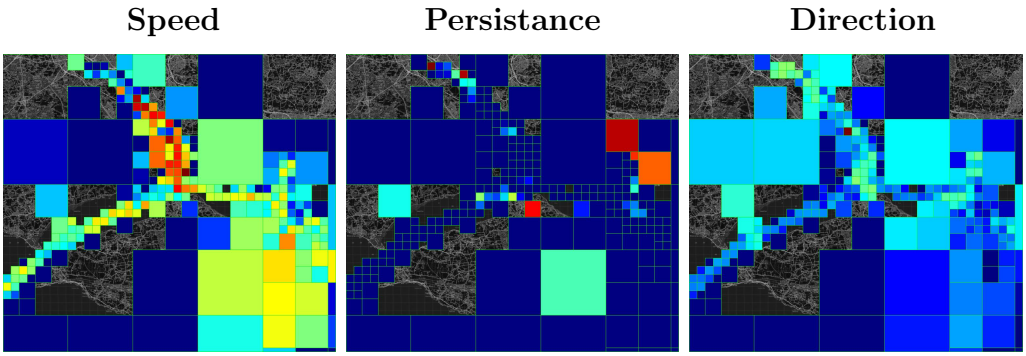


Figure 7.4: The entropy of distributions of speed (a), persistence (b), and direction (c) in the maritime 66 hour data set. Entropy gives an indication of how structured the region of the scene is. Low entropy indicates less constrained behaviour.

and there is a higher range of observed features upon the vertical. Persistence is very highly structured along the shipping routes vertically and horizontally. Only at ports, on the north side of the Isle of Wight, Portsmouth harbor, and Southampton harbor do we see much entropy in the persistence. The exception being on the east side of the Isle of Wight which acts as a waiting area for ships before preceding up Southampton Water. Directional entropy is low throughout the scene indicating well defined paths being taken with little overlap or crossing.

Our demonstration of scene context upon the maritime domain highlights clear structure within the maritime scene. We expected to see far more constrained motion within the maritime domain than the human surveillance environment; which we have observed. The increase in tracking accuracy will contribute a small amount to this finding. The algorithm for calculating human surveillance scene context generalises well to the maritime domain as there is direct analogy between human motion and ship motion, and both are modelled as point trajectories.

7.4.2 Social Context

To illustrate the generalisation of our social context algorithm to the maritime domain we next present examples of socially similar ships from the 66 hour AIS dataset we collected in house. The social estimation was calculated in batch mode, meaning at any point in time a social connection is considering past and future information from the 66 hour time span. For the sake of brevity we have illustrated only the top 5 social connections 7.5, and 2 counter examples 7.6.

With the removal of visual attention from the social model (there is no maritime analogy) the motion similarity is emphasised in the maritime domain. We witness that the top 5 social connections show strong motion similarity over a close time frame. We excluded 2 ships from the original top five as they were linked to one of the two ships and displayed very similar motion to those in image (b) 7.5. The ships in image (g) 7.5 are different to those in image (b) yet display very similar motion. The social similarity as applied to the maritime domain has successfully identified ships that intuitively appear to have a motion dependency. We did not however produce a ground truth for social connections, and thus not quantitative evaluation of performance has been provided. For comparison we illustrate 2 ships that display a low social similarity score 7.6. Image (a) indicates and example of near perfect

temporal overlap but dissimilar enough motion to lowly score the connection. In image (b), although there is a large degree of trajectory similarity the timing of the motions are dissimilar enough to cause a low connection score.

7.4.3 Anomalies

We next illustrate the use of our anomaly detection system upon the maritime behaviour dataset. As with the human surveillance the algorithm encodes the scene and social contextual information along with motion information. The QT coordinate system is used and Hierarchical clustering is used to populate a watchlist. The only distinction between the algorithm when applied to the maritime rather than the human domain is the removal of visual interest information.

We next illustrate the top anomalies found within the dataset. The anomalies are ranked by the same means as the human behaviour watchlist, Figure 7.7 and 7.7. For an impression of typical motion in the scene see Figure 7.1. The top anomaly 7.7 (a) found by our system displays a partial trajectory of a not uncommon behaviour (travelling between Isle of wight and Sounthampton harbour). However the AIS signal was very sparse. Due to the necessity to linearly interpolate the trajectory, the motion distribution is very peaked at the points of interpolation. We believe this causes the highly abnormal motion distribution. The second anomaly 7.7 (b) demonstrates an abnormal behaviour of leaving Southampton harbor and immediately looping back. Anomaly three 7.7 (c) shows a clearly abnormal short erratic motion across a traffic lane. Anomaly four 7.8 (d) shows an even starker erratic trajectory around the port area. The fifth anomaly 7.8 (e) shows an interesting case of sub-sampling time. The behaviour is that of entering the from the east and stopping in a port, which is seen often. However due to the time window this dataset was collected over the ship is seen for only part of the approach trajectory with a particularly sparse signal. The effect of this is to produce an abnormal motion distribution causing this normal behaviour to rank highly. We consider this to be a false positive. The final anomaly shown here 7.8 (f) shows part of a trajectory of a ship passing the Isle of Wight looping back and then going to harbor at the Isle of Wight. These six anomalies are only the top from the watch list produced. Our analysis of the remaining anomalies verifies that the trajectories ranked highly are particularly novel when compared to the remaining ships motion. Low ranked motions come from a set of a few common behaviours, typically moving in from the east or west, stopping at harbor and leaving again. Additionally there are a number of repeated ferry routes that rank low on the watchlist.

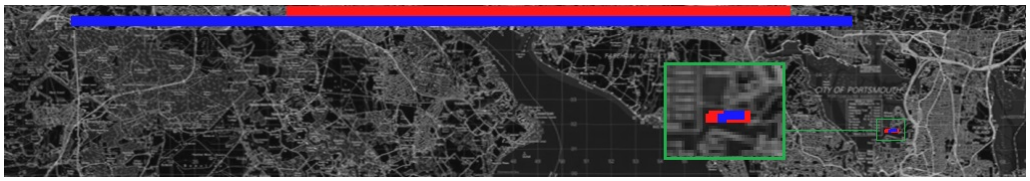
7.5 Conclusion

We find that the anomaly detection algorithm NN-RCO generalises well to the maritime domain. The anomalies that were highly ranked by our system show novel motion trajectories, thus verifying our approach on low noise motion data. Importantly this work validates that our approach is not domain specific. It does not fit a nuance of human behaviour in the environments we addressed, but instead it provides a more versatile approach towards behaviour outlier detection.

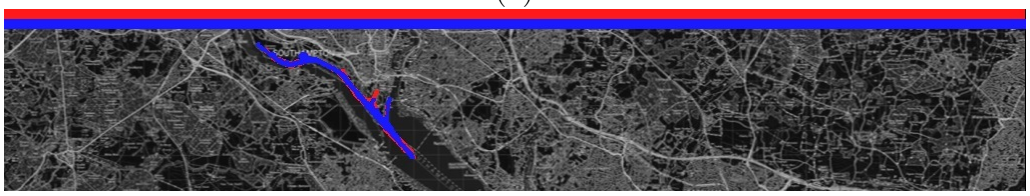
We note that partial tracks and sparse tracks are difficult for our system to handle. In the human surveillance domain trajectories are occluded for only short periods of time, or dropped entirely, and signals were not sparse. However, in the maritime domain a track may be lost when the ship stops broadcasting or the signal

is occluded, and then picked up again much later. This has the affect of warping the motion distribution which our system was not designed to handle, and as such causes a failure mode when the sparseness or occlusion is very high. Of the top six anomalies, it is our opinion that three (b), (c) and (d) are true positive anomalies, and three (a) (e) and (f) are produced by partial or sparse motion signals.

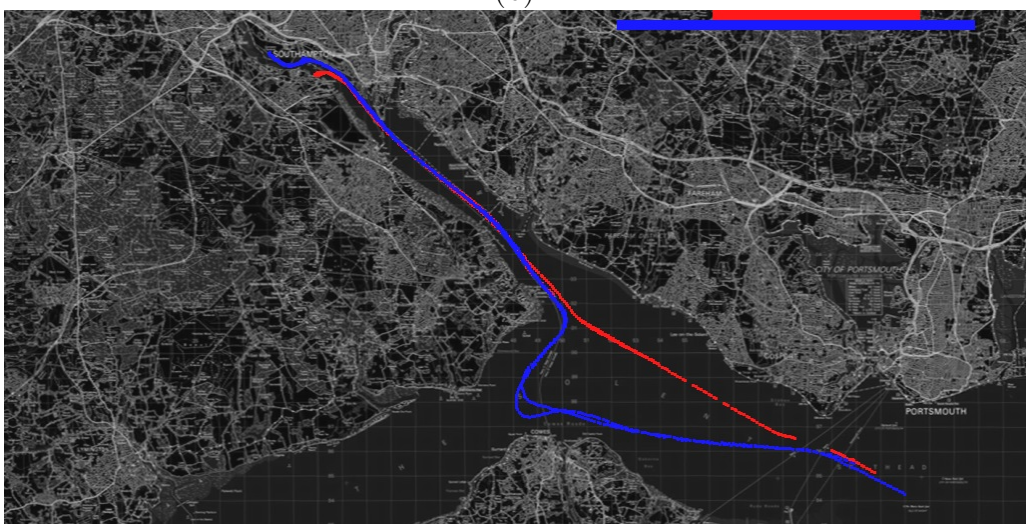
The implication of this chapter is that our NN-RCO algorithm generalises well to other motion domains. The algorithm could be applied to the air domain, or WAMI ground motion. To properly implement our algorithm upon wide area sea, ground, or air behaviour we would need to solve handling of partial and sparse tracks for the system to be competitive. We next conclude on all that has been presented in this work in the final chapter. We bring together all our work and evaluate our original hypothesis.



(a)

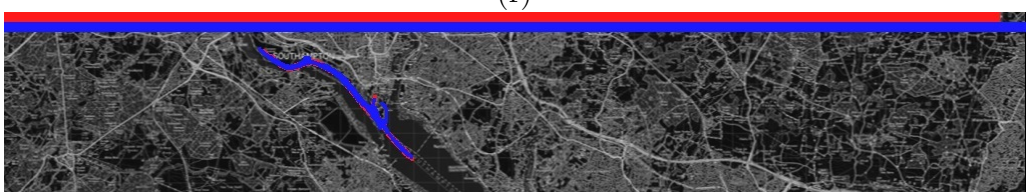


(b)



(e)

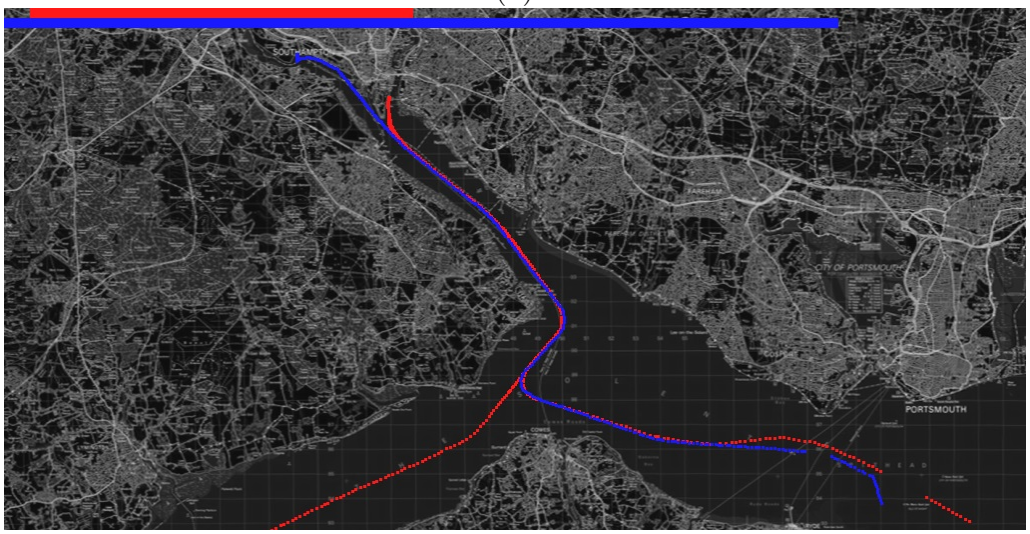
(f)



(g)

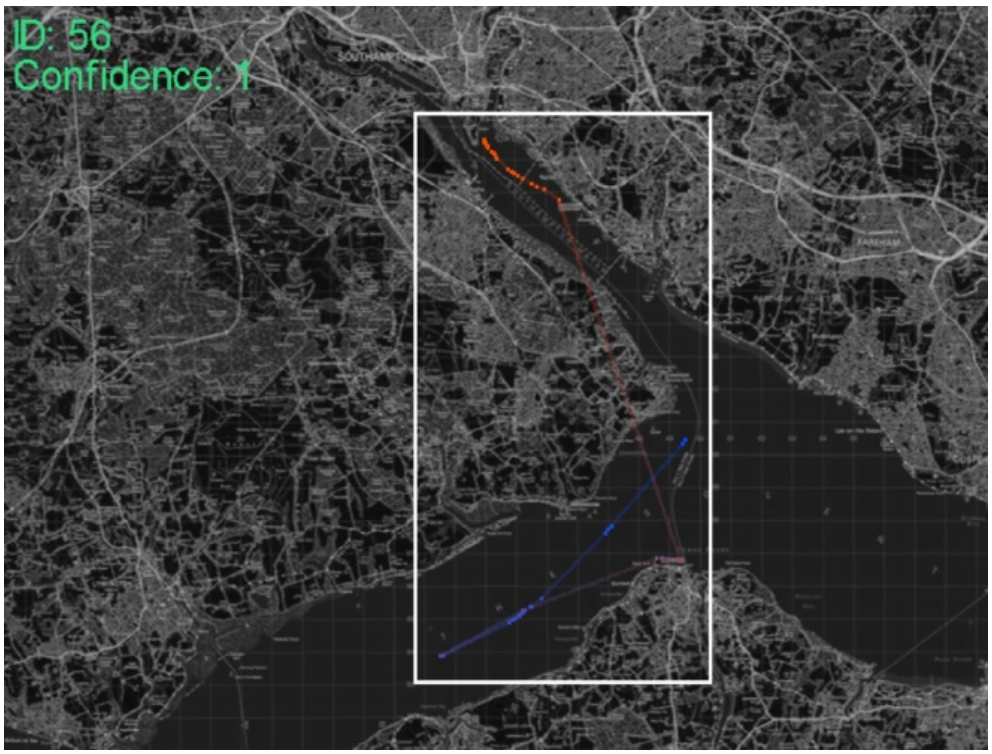
Figure 7.5: Top 5 connections found between ships. The trajectory of ship 1 is indicated by a blue track and the trajectory of ship 2 with a red track. Temporal overlap between two ships is indicated by correspondingly coloured bars at the top of the images, the full width of the image translates to the full time span of the dataset.

(a)

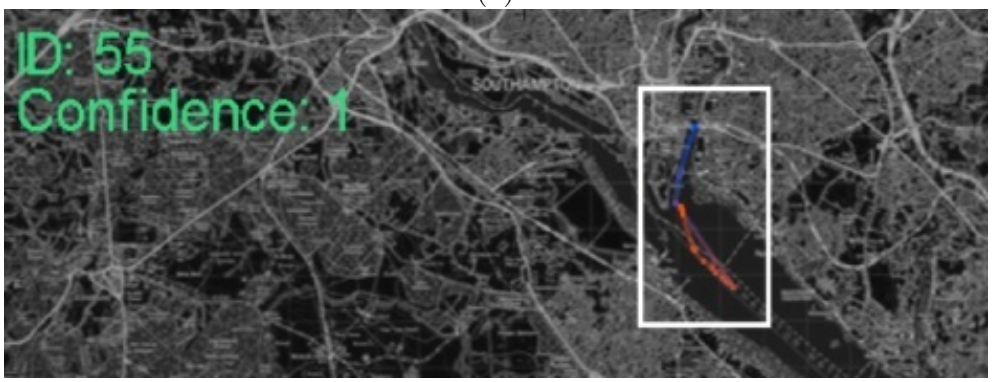


(b)

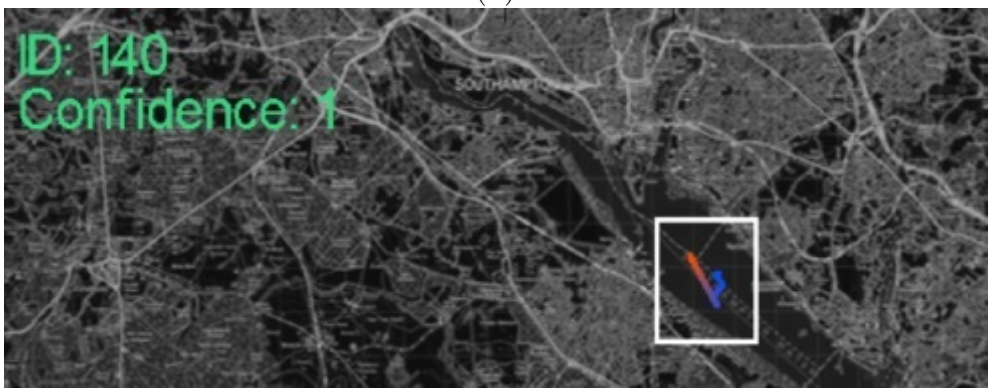
Figure 7.6: Examples of 2 true negative connections between ships. The trajectory of ship 1 is indicated by a blue track and the trajectory of ship 2 with a red track. Temporal overlap between two ships is indicated by correspondingly coloured bars at the top of the images, the full width of the image translates to the full time span of the dataset.



(a)

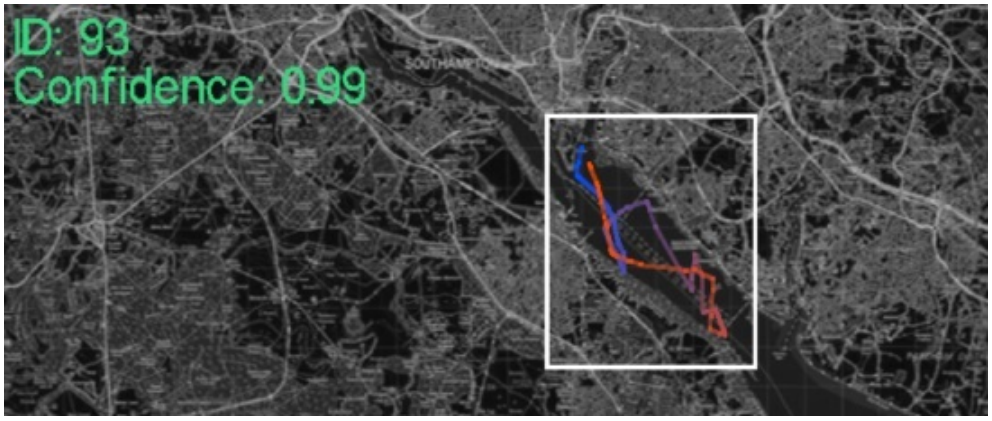


(b)

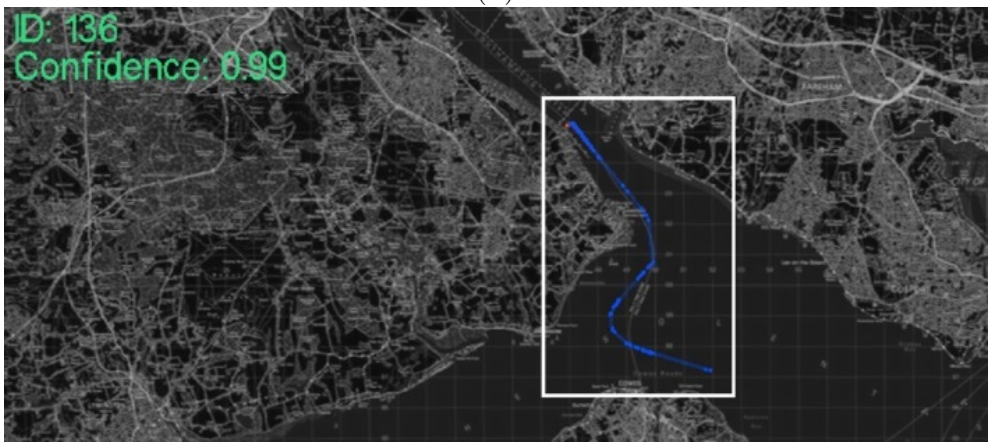


(c)

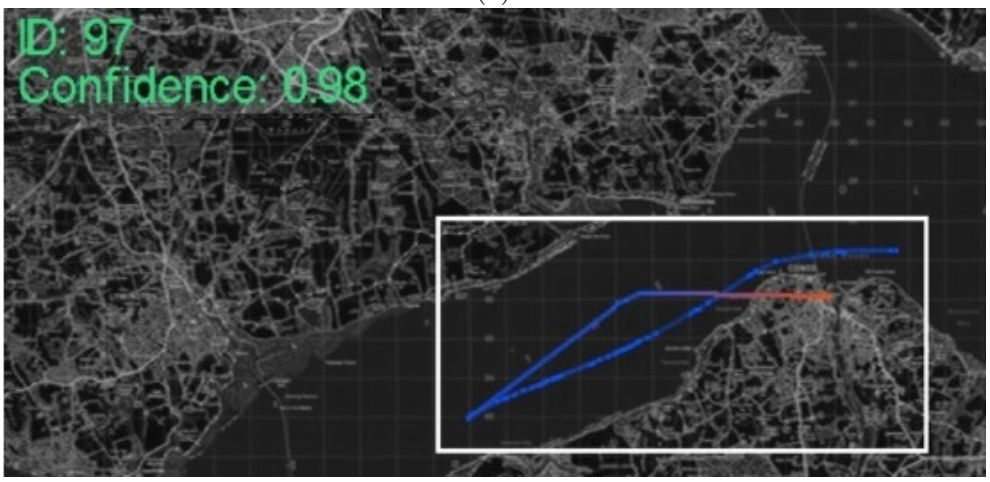
Figure 7.7: The top three ranked anomalies from the maritime dataset. The sequence of observations for the trajectory displayed is indicated by progression from blue to red. A white box aids locating the trajectory. The ship ID (1-199) and anomaly confidence is provided in the top left of each image.



(d)



(e)



(f)

Figure 7.8: The top 4th, 5th ,and 6th ranked anomalies from the maritime dataset. The sequence of observations for the trajectory displayed is indicated by progression from blue to red. A white box aids locating the trajectory. The ship ID (1-199) and anomaly confidence is provided in the top left of each image.

Chapter 8

Conclusion

The aim of this thesis was "to investigate existing theory and algorithms with the capability to detect abnormal human behaviour in surveillance or maritime data, evaluate opportunities to improve the existing capability of such techniques, and propose and evaluate algorithms to better detect abnormal behaviour". We now reiterate and evaluate how we have reached this aim. We bring together our work we presented in chapters 2 through to Chapter 7. We first provide a detailed list of our contributions and conclusions from each section of the thesis 8.1. We follow this by outlining the theory and algorithms that were only partially investigated or remain to be investigated in section 8.2. We present a list of applications our research has had in section 8.3 and we finish with a final conclusion and remarks in section 8.4.

8.1 Contributions

We first enumerate our contributions chronologically and relate them to our research objectives that we detailed in the introduction, Chapter 1, and background, Chapter 2, section of our work.

In the background chapter 2 we first presented the need for human behaviour anomaly detection in surveillance video. However, how this was achieved currently and what potential there was for more effective systems was not immediately answered. We contribute a review of relevant literature in behaviour modelling and anomaly detection. We found in particular two distinct approaches which dominate human behaviour anomaly detection in surveillance. The first defines behaviour as an agent activity and builds a human centric behaviour description. The second approach, non-human centric, seeks to define anomalies as patches of motion in the image stream. We concluded that human centric approaches have the advantage of encoding more information about the interaction of the agents responsible for behaviour, longer term profiling can be achieved, and tracking of humans can lead to further features such as head pose and contextual features being derived. The events in human surveillance, such as loitering, running, chasing, or queuing, are highly variable in appearance, motion, and context. Thus by their nature they lend themselves more to non-parametric representation. Non-parametric approaches are better adapted to the variation and dynamic appearance. Of particular interest are statistical techniques which implicitly model the segmentation of behaviour classes; such as nearest-neighbour. Our review of relevant literature, theory, and practices in anomaly detection revealed a number of areas for possible novel insight and algorithms. In particular we found there exists a gap in the state of the art when considering the implementation and analysis of automatic contextual infor-

mation in human surveillance. There was an opportunity to enhance and evaluate human behaviour anomaly detection using social modelling and scene understanding in surveillance. We identified the need for a method which implements and utilises contextual information about social connections to better classify abnormal behaviour in human surveillance. And we found there was scope to demonstrate the efficacy of scene modelling in human behaviour anomaly detection. A finding that drove much of our later work was the opportunity to propose, implement, and evaluate the use of head pose information in social modelling, scene modelling, and behaviour analysis.

Objective 1: Propose algorithms to deliver additive social context information into an anomaly detection system Following from our review of literature we are in the position to propose algorithms to utilise social and scene context information in a human behaviour anomaly detection system. In Chapter 4 we propose a social connection classification algorithm which uses properties of motion and the mutual information metric to identify the existence of social connections between pedestrians. The technique is largely based upon the appearance of a connection due to motion dependency between individuals. We use the weighted sum of speed, direction, proximity, and temporal overlap metrics to characterise the motion similarity. Speed and direction similarity are measured using mutual information, whilst proximity and temporal overlap use Euclidean distance. We test the social context classification against an independently constructed ground truth for social connections. Classifying social connections in the PETS 2007 data using parameters trained in the PETS 2006 data achieved a TPR of 0.92 and a FPR of 0.092, see Figure 4.3 (a). There are a greater number of false positive social connections in the Oxford data. The optimal result found 0.412 TPR and 0.0149 FPR. Our novel method draws from our findings from Chapter 2 to bring together the work of several other methods. We further enhance this method later using visual attention, see Chapter 5. This contribution has been published as part of [55].

Objective 2: Propose algorithms to deliver additive scene context information into an anomaly detection system We contribute a method of classifying regions of the scene in which the behaviour shows a dependency upon the local region. We classify 3 different regions (Idle region, traffic regions, divergence regions), where the lack of either three defaults to a 'general area'. We found well defined regions for the idle, divergence and traffic region in the PETS data which fit with the intuitive interpretation of the scene, see Figure 4.4. The Oxford data held well defined areas for the traffic region and the divergence region. However the idle region hardly featured. This finding fits with the highly structured nature of the Oxford data in which there are very few stationary tracks. As our approach is data driven, scene regions are defined by virtue of being a tool for segmenting the behaviour space rather than fitting an intuitive interpretation of scene regions. We later enhance the scene context algorithm, removing the need for a-priori region definitions and removing hard boundaries. This contribution has been published in [55].

Objective 3: Propose a novel algorithm for determining human behaviour anomalies which integrates contextual information into the analysis Our second biggest contribution is our evaluation of our proposed algorithm which proves the hypothesis that social and scene contextual information improves human behaviour anomaly detection.Our approach is unsupervised, and as such

anomalies are discovered due to their contrasting nature to previously observed behaviour. We distinguish between the *normality* of a behaviour and the *expectation* of a behaviour. The expectation of a behaviour is how likely it is to occur next, whereas the normality of a behaviour is how permitted the behaviour is in the scene; how legitimate it is. However, frequency-based anomaly detection suffers under the following assumption: that the normality of any observed behaviour is proportional to the relative frequency of observations of the behaviour. Whilst we can expect abnormal events to be rare, it is not the case that normal events are all frequent, and proportionally represented. A frequency-based analysis reveals expectation of each behaviour to occur next, not the intrinsic normality of the behaviour itself, thus missing the mark. Thus, contrary to the trend in contemporary work which focusses upon a frequency-based analysis to determine the normality of behaviour observations we utilise nearest neighbour clustering-based approach, where the degree to which something is an anomaly is based upon the distance in metric space to its nearest K-sized behaviour cluster. Thus, the onus is upon an effective metric in the behaviour space in which events are represented, see Chapter 4 for more details. The contribution here focusses upon the demonstration of the validity and effectiveness of our contextual information system. The main focus is upon the power of the features used (Scene context, social context, motion) rather than the anomaly detection algorithm itself. We implement our system upon 4 different datasets to test the validity. We find that in the three PETS-2007 datasets we observe that the addition of scene context improves the TPR over FPR detection of anomalies over all datasets in comparison to the no-context baseline. This is most significantly observed in Scene 04, Figure 4.6 (c). The significant result is that with the inclusion of both social context and scene context the TPR is improved above the TPR of scene context inclusion alone. This is due to the inclusion of the capability introduced by the social context to deny self-justifying groups and propagate anomalies within social groups. Particularly in PETS Scene 04, we observe that by propagating low likelihood scores throughout the group the bulk of true positive anomalies are discovered earlier, reducing the FPR from 0.2 to 0.03, see Figure 4.6 (c). The overall classification score with both social and scene context for all PETS-2007 data is shown in Figure 4.8. We later propose a second anomaly detection algorithm which includes contextual information, Chapter 6. This second system is an improved iteration of the current system. This contribution has been published as part of [55].

Objective 4: Demonstrate the entire pipeline of our proposed algorithm upon real world surveillance data We refer now to the second iteration of the anomaly detection algorithm we present in this thesis in Chapter 6. The main difference between the two systems are the introduction of hierarchical clustering to overcome thresholding and the introduction of head pose information. Secondarily the final version is more principled in its treatment of contextual information. The final anomaly detection algorithm, implementation, and evaluation is our main contribution of the thesis. It demonstrates upon real world surveillance data the effectiveness of contextual information in a novel system to detect subtle or hard to detect human behaviour anomalies. The feature extraction part of the pipeline that forms this contribution has been published as part of [55], [54], [7], and [8].

Objective 5: Demonstrate the feasibility and quantify the effectiveness of contextual information in human behaviour anomaly detection on real world data we evaluate our system upon PETS scene 4, PETS scene 0, and the

Oxford dataset. The Oxford data is different to the PETS data in that it presents a far simpler behaviour set. Scene segmentation is almost trivial in the Oxford data, however the social grouping is more complex due to the structured motion within the scene. The proof of the efficacy of our contextual work is provided in the evaluation of the behaviour analysis 6.7

Objective 6: Implement head pose estimation and utilise the information in our contextual work and behaviour analysis Head pose extraction is presented in the feature extraction chapter of the thesis 3. We follow the work of Benfold [13] with some minor alterations to improve accuracy at the cost of speed. We utilise the head pose information in the social group classification work and the scene segmentation work from Chapter 5. The main contribution we make is the use of visual attention in a social estimation and scene modelling. We hypothesised that socially connected individuals display this through the visual attention feature by either looking towards each other or correlating attention. We have validated this by looking specifically for these two cases and improving upon a purely motion-based social clustering. This contribution has been published as part of [54].

Objective 7: Evaluate our proposed algorithm upon real world surveillance data, demonstrate the efficacy of our approach and assess our algorithms in light of other state of the art approaches We contribute a detailed qualitative evaluation of our method against other state of the art approaches in section 6.7.4.

Objective 8: Evaluate and quantify the effectiveness of head pose information in contextual information sources Our approach successfully classifies social groups in the scene, achieving a true positive rate of 0.93 - 0.95 at a false positive rate of 0.02 - 0.05, depending on the dataset, and using ground truth visual attention cues. Using fully automatic feature extraction we achieved a true positive rate of 0.88 - 0.92 at a false positive rate of 0.04 - 0.07. This finding, and demonstration, opens a new methodology for automatic social estimation which may have implication beyond security; it may feature in marketing and crowd control analysis. This contribution has been published as part of [54].

Objective 9: Evaluate and quantify the effectiveness of head pose information in human behaviour anomaly detection We demonstrate the impact of introducing visual attention upon the PETS scene 4 data. We witness a higher ranking of the first two ranked anomalies, and a large improvement in detection of the fourth and sixth highest ranked anomalies in the data. Analysis of the output shows the system ranks the anomalies in the same order, but with the inclusion of fewer false positives when visual attention is included. We demonstrate here 6.10 that visual attention has a beneficial role upon anomaly detection in this dataset. Thus we validate the intuition that abnormal behaviour is often partially characterised by an abnormal visual attention in the scene. We find that noise in the data generated from automatic head pose estimation and social modelling severely impacts the Scene 00 analysis however has little or no impact on the Scene 04 and Oxford data, even though the level of noise was similar for all scenes.

Drawing from our assessment of having met the above objectives above we can evaluate whether our thesis was found to be true or false. We stated in the intro-

duction of this work, Chapter 1, that our hypothesis was:

> *Feature rich, data driven anomaly detection algorithms can remove the need for data intensive machine learning and expensive modelling techniques. By using contextual, motion, and head pose information we can separate heterogeneous behaviour clusters by increasing the interclass distance or reducing the intraclass distances, thus making outliers more salient. This allows for anomalies to be detected via the means of outlier detection.*

The hypothesis breaks into two main parts; by using contextual, motion, and head pose information we can make outliers more salient, and following from this, anomalies can be detected via the means of outlier detection. The detection of anomalies via outlier detection is a proven hypothesis already in the field, so the interpretation of the hypothesis is such that anomalies can be better detected using contextual information, given an outlier detection method. We implemented a feature rich method, using head pose information, contextual information, and motion information, where most commonly only motion is used. By doing this we avoided the need for a data intensive machine learning algorithm, such as the various nuanced HMMs. Instead we use a form of nearest neighbour clustering and hierarchical clustering to draw out outliers. Thus, we have satisfied the hypothesis that "Feature rich, data driven anomaly detection algorithms can remove the need for data intensive machine learning and expensive modelling techniques". We demonstrate extensively in Chapter 4 that contextual information can be used to improve the detection of anomalies, and in Chapter 6 that head pose information and contextual information increases the outlier score of anomalies. We applied outlier detection in the form of hierarchical clustering to populate a ranked outlier list of nearest-neighbour-clusters such that any improvement in anomaly detection necessitates that the outlier score of anomalies are greater, entailing that our method has increased interclass distance or reduced the intraclass distances with respect to the anomaly behaviours. Following from the experiments that prove context and head pose information to be effective means of teasing out outliers we can conclude we have satisfied the hypothesis that "By using contextual, motion, and head pose information we can separate heterogeneous behaviour clusters by increasing the interclass distance or reducing the intraclass distances, thus making outliers more salient". We consider our thesis to have been found true from this body of work we present.

8.2 Future Work

We next investigate potential future work based upon the findings and conclusions of our work.

Minimum description length social classification. We extended the social similarity metric from Chapter 5 to include classification of social clusters based upon the principle of Minimum Description length (MDL). However, we did not include the findings as our research extended as far as a proof of principle on a single dataset. The concept is such that the social similarity matrix, which details the extent of the appearance of a social connection between every two pairings of people, encodes all possible hypothesis of social groupings. A social grouping hypothesis is a particular classification of individuals into social group identities. Merely applying a threshold to the social similarity matrix does not cluster connections into

groups, but merely states which connections we wish to consider. Using MDL we can associate potentially connected individuals into groups by selecting the optimal hypothesis which encodes all the connections most efficiently. We initially do this by using the one-to-one correspondence between code length functions and probability distributions to iteratively search for the best hypothesis. We start assuming that no two people are in a social group, and calculate the description length of encoding this hypothesis. We then propose a new hypothesis identical to the previous but with a single social group consisting of the highest connection from the social similarity matrix. We iteratively add new individuals to groups until we converge upon the most compact encoding all of connections into groups. Our initial results are promising showing better classification of groups than taking the optimal threshold from the social connection matrix, meaning that this method is capable of adding information by assessing the 2nd degree connections (A-B, B-C, therefore A-B-C).

Crowd-sourced anomaly detection. We propose the idea of using the people within a surveillance scene as the source of anomaly detection for security applications. By tracking the head pose of people within a scene, and modelling the ambient motion, it should be possible to detect when there is an abnormally high and sudden interest in a region of the scene. Such events that draw an abnormal amount of attention are synonymous with events of interest to security staff and as such should be flagged. With such a technique it would in fact be possible to detect events happening outside the field of view of the sensor. The convergence of visual attention would indicate the location of the event.

Empirical evaluation of scene context. We were unable, within the scope of our research, to carry out an empirical validation of scene regions for the scene context. It may be possible to ground truth scene structure using synthetic data or real world controlled data.

Deep learning head pose estimation. The recent surge in deep learning techniques has presented many methods particularly capable at classification tasks. The variability in appearances for any head pose classes and the abundance of training data lends itself particularly to deep learning techniques. We have shown that improved head pose estimation increases the efficacy of social grouping estimation through an analysis of the impact of feature noise, see Figure 5.4. Thus, we propose using a Deep Belief Network to improve head pose classification results and consequently social clustering.

8.3 Applications

We next enumerate the applications where we have used our research in industry. Details are withheld where naming particulars of the application is restricted.

The primary use of our research has been in a Maritime behaviour analysis project. This project aims to improve maritime situational awareness for large assets using a mixture of radar and AIS information. Our system is used to monitor the movement of other ships and small crafts to determine suspicious behaviour or threatening behaviour that should be brought to the attention of the operator. Our behaviour analysis method forms the long term behaviour analysis in combination with a faster short term motion abnormality detector which specialises in detecting acute motion anomalies. Our method applies the social model work to the maritime domain in order to detect groups of ships such as fishing fleets, convoys, and tugs. The importance of this step is that it is used to detect when a member of a group suddenly stops acting like the rest of the group, as this may be indicative of a ship

hiding amongst legitimate behaviour.

The human detection and tracking system we built for this thesis has been used as a means of detecting and tracking people through a small mounted head ups display for military use, similar to the kind that may be attached to a firearm. For this challenge we had to make the detection and tracking very lightweight to reduce the latency of the tracking location to a minimum.

The detection and tracking has been combined with re-identification algorithms in order to carry out prolonged surveillance tasks in which the identity of people coming and going is of importance. However more details cannot be given about this project.

8.4 Final Remarks

We proposed the NN-RCO algorithm as a means of encoding contextual information into an anomaly detection framework. Our method is not domain specific; it can be generalised to other types of data. Contextual information is used as a means of increasing interclass distance and reducing intraclass distance, so as long as a relevant contextual feature can be extracted, derived, or estimated it can be used to enhance anomaly detection for your domain. Given we have shown the power of data driven contextual information in the human and maritime surveillance tasks we hope to see further adoption of contextual information in situational and surveillance tasks in research and industry. Furthermore, we demonstrate that visual attention estimation based upon head pose can be used to improve behaviour analysis and social group estimation. This finding paves the way for interesting future techniques extending behaviour analysis beyond motion alone. The main point that we wish people to take from our research is that feature rich, data driven anomaly detection algorithms such as NN-RCO can remove the need for data intensive machine learning and expensive modelling techniques by creating further separation between representations of behaviours allowing the classification of more subtle outlier behaviours.

Bibliography

[1] Caviar dataset: http://homepages.inf.ed.ac.uk/rbf/caviar/, 2005.

[2] Pets2007 dataset: www.cvg.rdg.ac.uk/pets2007/data.html, 2007.

[3] T. Ahmed, B. Oreshkin, and M. Coates. Machine learning approaches to network anomaly detection. *SysML Workshop*, 2007.

[4] K. R. T. Aires, A. M. Santana, and A. A. D. Medeiros. Optical flow using color information. In *Proceedings of the 2008 ACM symposium on Applied computing*. ACM Press, 2008.

[5] B. Antic and B. Ommer. Video parsing for abnormality detection. In *2011 International Conference on Computer Vision*. IEEE, Nov 2011.

[6] S. O. Ba and J. Odobez. Multiperson visual focus of attention from head pose and meeting contextual cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):101–116, Jan 2011.

[7] R. H. Baxter, M. Leach, and N. M. Robertson. Tracking with intent. In *2014 Sensor Signal Processing for Defence (SSPD)*. IEEE, Sep 2014.

[8] R. H. Baxter, M. J. V. Leach, S. S. Mukherjee, and N. M. Robertson. An adaptive motion model for person tracking with instantaneous head-pose features. *IEEE Signal Process. Lett.*, 22(5):578–582, May 2015.

[9] L. Bazzani, M. Cristani, D. Tosato, M. Farenzena, G. Paggetti, G. Menegaz, and V. Murino. Social interactions by visual focus of attention in a three-dimensional environment. *Expert Systems*, 30(2):115–127, Jul 2012.

[10] B. Benfold. The acquisition of coarse gaze estimates in visual surveillance. *Thesis, Robotics Research Group, University of Oxford*, 2011.

[11] B. Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video. In *CVPR 2011*. IEEE, Jun 2011.

[12] B. Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video, 2011. Data.

[13] B. Benfold and I. Reid. Unsupervised learning of a scene-specific coarse gaze estimator. In *2011 International Conference on Computer Vision*. IEEE, Nov 2011.

[14] P. Bilinski, F. Bremond, and M. Kaaniche. Multiple object tracking with occlusions using HOG descriptors and multi resolution images. In *3rd International Conference on Imaging for Crime Detection and Prevention (ICDP 2009)*. IET, 2009.

[15] N. Bomberger, B. Rhodes, M. Seibert, and A. Waxman. Associative learning of vessel motion patterns for maritime situation awareness. In *2006 9th International Conference on Information Fusion*. IEEE, Jul 2006.

[16] P. Bradley, U. Fayyad, and C. Reina. Scaling clustering algorithms to large databases. *International Conference on Knowledge Discovery and Data Mining*, 4, 1998.

[17] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection. *CSUR*, 41(3):1–58, Jul 2009.

[18] D. Chau, F. Bremond, and M. Thonnat. A multi-feature tracking algorithm enabling adaptation to context variations. In *4th International Conference on Imaging for Crime Detection and Prevention 2011 (ICDP 2011)*. IET, 2011.

[19] D. P. Chau, F. Bremond, and M. Thonnat. Object tracking in videos: Approaches and issues. *The International Workshop, Rencontres UNS-UD*, 2013.

[20] C. Chen and J. Odobez. We are not contortionists: Coupled adaptive learning for head and body orientation estimation in surveillance video. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Jun 2012.

[21] N. G. Chitaliya and A. I. Trivedi. Novel block matching algorithm using predictive motion vector for video object tracking based on color histogram. In *2011 3rd International Conference on Electronics Computer Technology*. IEEE, Apr 2011.

[22] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*. IEEE Comput. Soc, 2000.

[23] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Trans. Pattern Anal. Machine Intell.*, 25(5):564–577, May 2003.

[24] M. Cristani, R. Raghavendra, A. D. Bue, and V. Murino. Human behavior analysis in video surveillance: A social signal processing perspective. *Neurocomputing*, 100:86–97, Jan 2013.

[25] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2005.

[26] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 1977.

[27] W. Eberle and L. Holder. Incremental anomaly detection in graphs. In *2013 IEEE 13th International Conference on Data Mining Workshops*. IEEE, Dec 2013.

[28] J. Edlund, M. Grnkvist, A. Lingvall, and E. Sviestins. Rule-based situation assessment for sea surveillance. In B. V. Dasarathy, editor, *Multisensor, Multisource Information Fusion: Architectures, Algorithms, and Applications 2006*. SPIE, Apr 2006.

[29] M. Ester, H.-P. Kriegel, J. Sanger, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of Second International Conference on Knowledge Discovery and Data Mining*, 1996.

[30] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, Sep 2010.

[31] F. Fooladvandi, C. Brax, P. Gustavsson, and M. Fredin. Signature-based activity detection based on bayesian networks acquired from expert knowledge. , 2009.

[32] W. Ge, R. T. Collins, and B. Ruback. Automatically detecting the small group structure of a crowd. In *2009 Workshop on Applications of Computer Vision (WACV)*. IEEE, Dec 2009.

[33] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Jun 2014.

[34] N. Gourier, J. Maisonnasse, D. Hall, and J. L. Crowley. Head pose estimation on low resolution images. In *Multimodal Technologies for Perception of Humans*, pages 270–280. Springer Science Business Media, 2007.

[35] H. Hajji. Statistical analysis of network traffic for adaptive faults detection. *IEEE Trans. Neural Netw.*, 16(5):1053–1063, Sep 2005.

[36] Z. He, X. Xu, and S. Deng. Discovering cluster based local outliers. *Pattern Recognition Letters*, 2003.

[37] T. M. Hospedales, J. Li, S. Gong, and T. Xiang. Identifying rare and subtle behaviors: A weakly supervised joint topic model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2451–2464, Dec 2011.

[38] W. Hu, D. Xie, T. Tan, and S. Maybank. Learning activity patterns using fuzzy self-organizing neural network. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, 34(3):1618–1626, Jun 2004.

[39] X. Hu, S. Hu, X. Zhang, H. Zhang, and L. Luo. Anomaly detection based on local nearest neighbor distance descriptor in crowded scenes. *The Scientific World Journal*, 2014:1–12, 2014.

[40] Y. Huang and I. Essa. Tracking multiple objects through occlusions. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2005.

[41] F. Johansson and G. Falkman. Detection of vessel anomalies - a bayesian network approach. In *2007 3rd International Conference on Intelligent Sensors, Sensor Networks and Information*. IEEE, 2007.

[42] Z. Kalal, J. Matas, and K. Mikolajczyk. Online learning of robust object detectors during unstable tracking. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*. IEEE, Sep 2009.

[43] Z. Kalal, J. Matas, and K. Mikolajczyk. P-n learning: Bootstrapping binary classifiers by structural constraints. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, Jun 2010.

[44] Z. Kalal, K. Mikolajczyk, and J. Matas. Face-TLD: Tracking-learning-detection applied to faces. In *2010 IEEE International Conference on Image Processing*. IEEE, Sep 2010.

[45] Z. Kalal, K. Mikolajczyk, and J. Matas. Forward-backward error: Automatic detection of tracking failures. In *2010 20th International Conference on Pattern Recognition*. IEEE, Aug 2010.

[46] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-learning-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1409–1422, Jul 2012.

[47] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35, 1960.

[48] T. Kohonen. Self-organising maps. *Springer 3rd Edition*, 2000.

[49] N. Krahnstoever, M.-C. Chang, and W. Ge. Gaze and body pose estimation from a distance. In *2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, Aug 2011.

[50] J. B. Kraiman, S. L. Arouh, and M. L. Webb. Automated anomaly detection processor. , 2002.

[51] P. S. Kumar, P. Guha, and A. Mukerjee. Colour and feature based multiple object tracking under heavy occlusions. In *Advances in Pattern Recognition - Proceedings of the Sixth International Conference*. World Scientific Publishing Co. Pte. Ltd., 2007.

[52] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori. Discriminative latent models for recognizing contextual group activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(8):1549–1562, Aug 2012.

[53] R. Laxhammar and G. Falkman. Inductive conformal anomaly detection for sequential detection of anomalous sub-trajectories. *Annals of Mathematics and Artificial Intelligence*, Sep 2013.

[54] M. Leach, R. Baxter, N. Robertson, and E. Sparks. Detecting social groups in crowded surveillance videos using visual attention. In *2014 IEEE Computer Vision and Pattern Recognition Workshops*. IEEE, Jun 2014.

[55] M. J. Leach, E. Sparks, and N. M. Robertson. Contextual anomaly detection in crowded surveillance scenes. *Pattern Recognition Letters*, 44:71–79, Jul 2014.

[56] M. J. Leach, E. Sparks, and N. M. Robertson. Contextual anomaly detection in crowded surveillance scenes. *Pattern Recognition Letters*, 44:71–79, Jul 2014.

[57] W. Lee and D. Xiang. Information-theoretic measures for anomaly detection. In *Proceedings 2001 IEEE Symposium on Security and Privacy. S&P 2001*. IEEE Comput. Soc, 2001.

[58] V. K. Levent Ertoz, Michael Steinbeck. A new shared nearest neighbour clustering algorithm and it's applications. *SIAM Workshop on Clustering High Dimensional Data and it's Applications*, 2002.

[59] J. Li, S. Gong, and T. Xiang. Scene segmentation for behaviour correlation. In *Computer Vision – ECCV 2008*, pages 383–395. Springer Science Business Media, 2008.

[60] C. C. Loy. *Activity understanding and Unusual Event Detection in Surveillance Videos*. PhD thesis, 2010.

[61] C. C. Loy, T. Xiang, and S. Gong. Detecting and discriminating behavioural anomalies. *Pattern Recognition*, 44(1):117–132, Jan 2011.

[62] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of Imaging Understanding Workshop*, 1981.

[63] D. Makris and T. Ellis. Learning semantic scene models from observing activity in visual surveillance. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, 35(3):397–408, Jun 2005.

[64] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Jun 2009.

[65] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Jun 2014.

[66] N. Oliver, B. Rosario, and A. Pentland. Statistical modelling of human interactions. *CVPR Workshop on Interpretation of Visual Motion*, 1998.

[67] J. Pan and B. Hu. Robust occlusion handling in object tracking. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Jun 2007.

[68] M. Panda and M. R. Patra. Network intrusion detection using naive bayes. *IJCSNS International Journal of Computer Science and Network Security*, 2007.

[69] V. Papadourakis and A. Argyros. Multiple objects tracking in the presence of long-term occlusions. *Computer Vision and Image Understanding*, 114(7):835–846, Jul 2010.

[70] S. Pellegrini, A. Ess, K. Schindler, and L. V. Goo. You'll never walk alone: Modeling social behavior for multi-target tracking. *IEEE 12th International Conference*, 2009.

[71] M. Ranzato, Y.-L. Boureau, and Y. LeCun. Sparse feature learning for deep belief networks. In *Neural Information Processing Systems (NIPS)*, 2007.

[72] B. Rhodes, N. Bomberger, M. Seibert, and A. Waxman. Maritime situation monitoring and awareness using learning mechanisms. In *MILCOM 2005 - 2005 IEEE Military Communications Conference*. IEEE, 2005.

[73] B. Ristic, B. L. Scala, M. Morelande, and N. Gordon. Statistical analysis of motion patterns in ais data: Anomaly detection and motion prediction. , 2008.

[74] N. M. Robertson and I. D. Reid. Automatic reasoning about causal events in surveillance video. *EURASIP Journal on Image and Video Processing*, 2011:1–19, 2011.

[75] M. Rodriguez, S. Ali, and T. Kanade. Tracking in unstructured crowded scenes. In *2009 IEEE 12th International Conference on Computer Vision*. IEEE, Sep 2009.

[76] M. J. Roshtkhari and M. D. Levine. Online dominant and anomalous behavior detection in videos. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Jun 2013.

[77] X. Song, M. Wu, C. Jermaine, and S. Ranka. Conditional anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 19(5):631–645, May 2007.

[78] C. D. Stefano, C. Sansone, and M. Vento. To reject or not to reject: that is the question-an answer in case of neural classifiers. *IEEE Trans. Syst., Man, Cybern. C*, 30(1):84–94, 2000.

[79] M. Steinbach, P.-N. Tan, and V. Kumar. Introduction to data mining. *Addison-Wesley, Boston*, ISBN 0-321-32136-7, 2006.

[80] R. Stiefelhagen. Tracking focus of attention in meetings. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*. IEEE Comput. Soc, 2002.

[81] K. S. Sudipto Guha, Rajeev Rastogi. Rock: A robust clustering algorithm for categorical attributes. *15th IEEE International Conference on Data Engineering*, 2000.

[82] J. Sun, Y. Xie, H. Zhang, and C. Faloutsos. Less is more: Compact matrix decomposition for large sparse graphs. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pages 366–377. Society for Industrial and Applied Mathematics, Apr 2007.

[83] M. H. Tun, G. S. Chambers, T. Tan, and T. Ly. Maritime port intelligence using ais data. 2007.

[84] Š. Urban, M. Jakob, and M. Pěchouček. Probabilistic modeling of mobile agents' trajectories. In *Lecture Notes in Computer Science*, pages 59–70. Springer Science Business Media, 2010.

[85] X. Wang, K. T. Ma, G.-W. Ng, and W. E. L. Grimson. Trajectory analysis and semantic region modeling using a nonparametric bayesian model. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Jun 2008.

[86] V. R. V. Wenjie Hu, Yihua Liao. Robust anomaly detection using support vector machines. *International Conference on Machine Learning and ApplicationComputer Security*, 2003.

[87] J. Yan, Z. Lei, L. Wen, and S. Z. Li. The fastest deformable part model for object detection. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Jun 2014.

[88] T. Yu, S.-N. Lim, K. Patwardhan, and N. Krahnstoever. Monitoring, recognizing and discovering social networks. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Jun 2009.

[89] J. Zhu, O. Javed, J. Liu, Q. Yu, H. Cheng, and H. Sawhney. Pedestrian detection in low-resolution imagery by learning multi-scale intrinsic motion structures (MIMS). In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Jun 2014.